

SESSION
COMPUTER VISION AND APPLICATIONS

Chair(s)

TBA

An Algorithm for Mobile Vision-Based Localization of Skewed Nutrition Labels that Maximizes Specificity

Vladimir Kulyukin
Department of Computer Science
Utah State University
Logan, UT, USA
vladimir.kulyukin@usu.edu

Christopher Blay
Department of Computer Science
Utah State University
Logan, UT, USA
chris.b.blay@gmail.com

Abstract—An algorithm is presented for mobile vision-based localization of skewed nutrition labels on grocery packages that maximizes specificity, i.e., the percentage of true negative matches out of all possible negative matches. The algorithm works on frames captured from the smartphone camera's video stream and localizes nutrition labels skewed up to 35-40 degrees in either direction from the vertical axis of the captured frame. The algorithm uses three image processing methods: edge detection, line detection, and corner detection. The algorithm targets medium- to high-end mobile devices with single or quad-core ARM systems. Since cameras on these devices capture several frames per second, the algorithm is designed to minimize false positives rather than maximize true ones, because, at such frequent frame capture rates, it is far more important for the overall performance to minimize the processing time per frame. The algorithm is implemented on the Google Nexus 7 Android 4.3 smartphone. Evaluation was done on 378 frames, of which 266 contained NLs and 112 did not. The algorithm's performance, current limitations, and possible improvements are analyzed and discussed.

Keywords—computer vision; nutrition label localization; mobile computing; text spotting; nutrition management

I. Introduction

Many nutritionists and dieticians consider proactive nutrition management to be a key factor in reducing and controlling cancer, diabetes, and other illnesses related to or caused by mismanaged or inadequate diets. According to the U.S. Department of Agriculture, U.S. residents have increased their caloric intake by 523 calories per day since 1970. Mismanaged diets are estimated to account for 30-35 percent of cancer cases [1]. A leading cause of mortality in men is prostate cancer. A leading cause of mortality in women is breast cancer. Approximately 47,000,000 U.S. residents have metabolic syndrome and diabetes. Diabetes in children appears to be closely related to increasing obesity levels. The current prevalence of diabetes in the world is estimated to be at 2.8 percent [2]. It is expected that by 2030 the diabetes prevalence number will reach 4.4 percent. Some long-term complications of diabetes are blindness, kidney failure, and amputations. Nutrition labels (NLs) remain the main source of nutritional information on product packages [3, 4]. Therefore, enabling customers to use computer vision on their smartphones will likely result in

a greater consumer awareness of the caloric and nutritional content of purchased grocery products.



Figure 1. Skewed NL with vertical axis

In our previous research, we developed a vision-based localization algorithm for horizontally or vertically aligned NLs on smartphones [5]. The new algorithm, presented in this paper, improves our previous algorithm in that it handles not only aligned NLs but also those that are skewed up to 35-40 degrees from the vertical axis of the captured frame. Figure 1 shows an example of such a skewed NL with the vertical axis of the captured frame denoted by a white line. Another improvement designed and implemented in the new algorithm is the rapid detection of the presence of an NL in each frame, which improves the run time, because the new algorithm fails fast and proceeds to the next frame from the video stream.

The new algorithm targets medium- to high-end mobile devices with single or quad-core ARM systems. Since cameras on these devices capture several frames per second, the algorithm is designed to minimize false positives rather than maximize true ones, because, at such frequent frame capture rates, it is far more important to minimize the processing time per frame.

The remainder of our paper is organized as follows. In Section II, we present our previous work on accessible shopping and nutrition management to give the reader a broader context of the research and development presented in this paper. In Section III, we outline the details of our algorithm. In Section IV, we present the experiments with our algorithm and discuss our results. In Section V, we present our conclusions and outline several directions for future work.

II. Previous Work

In 2006, our laboratory began to work on ShopTalk, a wearable system for independent blind supermarket shopping [6]. In 2008 - 2009, ShopTalk was ported to the Nokia E70 smartphone connected to a Bluetooth barcode pencil scanner [7]. In 2010, we began our work on computer vision techniques for eyes-free barcode scanning [8]. In 2013, we published several algorithms for localizing skewed barcodes as well as horizontally or vertically aligned NLs [5, 9]. The algorithm presented in this paper improves the previous NL localization algorithm by relaxing the NL alignment constraint for up to 35 to 40 degrees in either direction from the vertical orientation axis of the captured frame.

Modern nutrition management system designers and developers assume that users understand how to collect nutritional data and can be triggered into data collection with digital prompts (e.g., email or SMS). Such systems often underperform, because many users find it difficult to integrate nutrition data collection into their daily activities due to lack of time, motivation, or training. Eventually they turn off or ignore digital stimuli [10].

To overcome these challenges, in 2012 we began to develop a Persuasive NUTrition Management System (PNUTS) [5]. PNUTS seeks to shift current research and clinical practices in nutrition management toward persuasion, automated nutritional information extraction and processing, and context-sensitive nutrition decision support. PNUTS is based on a nutrition management approach inspired by the Fogg Behavior Model (FBM) [10], which states that motivation alone is insufficient to stimulate target behaviors. Even a motivated user must have both the ability to execute a behavior and a trigger to engage in that behavior at an appropriate place or time.

Another frequent assumption, which is not always accurate, is that consumers and patients are either more skilled than they actually are or that they can be quickly trained to obtain the required skills. Since training is difficult and time consuming, a more promising path is to make target behaviors easier and more intuitive to execute for the average smartphone user. Vision-based extraction of nutritional information from NLs on product packages is a fundamental step in making proactive nutrition management easier and more intuitive, because it improves the user's ability to engage into the target behavior of collecting and processing nutritional data.

III. Skewed NL Localization Algorithm

A. Detection of Edges, Lines, and Corners

Our NL detection algorithm uses three image processing methods: edge detection, line detection, and corner detection. Edge detection transforms images into bitmaps where every pixel is classified as belonging or not belonging to an edge. The algorithm uses the Canny edge detector (CED) [11]. After the edges are detected (see Fig. 2), the image is processed with the Hough Transform (HT) [12] to detect lines (see Fig. 3). The HT algorithm finds paths in images that follow generalized polynomials in the polar coordinate space.



Figure 2. Original NL (left); NL with edges (right)

Corner detection is done primarily for text spotting because text segments tend to contain many distinct corners. Thus, image segments with higher concentrations of corners are likely to contain text. Corners are detected with the dilate-erode method [13] (see Fig. 4). Two stages of the dilate-erode method with different 5x5 kernels are applied. Two stages of dilate-erode with different kernels are applied. The first stage uses a 5x5 *cross* dilate kernel for horizontal and vertical expansions. It then uses a 5x5 *diamond* erode kernel for diagonal shrinking. The resulting image is compared with the original and those pixels which are in the corner of an aligned rectangle are found.

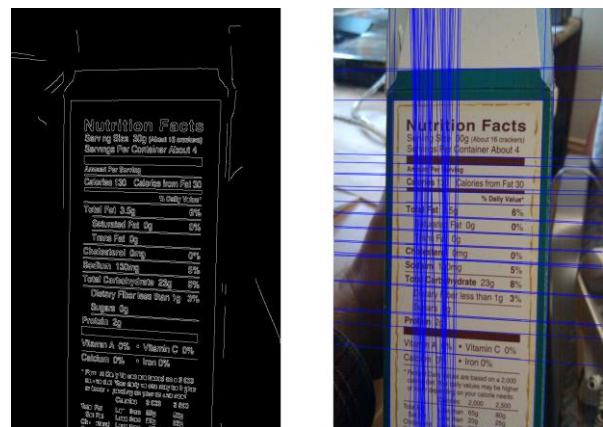


Figure 3. NL with edges (left); NL with lines (right)

The second stage uses a 5x5 X-shape dilate kernel to expand in the two diagonal directions. A 5x5 square kernel is used next to erode the image and to shrink it horizontally and vertically. The resulting image is compared with the original and those pixels which are in a 45 degree corner are identified. The resulting corners from both steps are combined into a final set of detected corners.

In Fig. 4, the top sequence of images corresponds to stage one when the *cross* and *diamond* kernels are used to detect aligned corners. The bottom sequence of images corresponds to stage two when the X-shape and *square* kernels are used to detect 45 degree corners. Step one shows the original input of each stage, step two is the image after dilation, step three is the image after erosion, and step four is the difference between the original and eroded versions. The resulting corners are outlined in red in each step to provide a basis of how the dilate-erode operations modify the input.

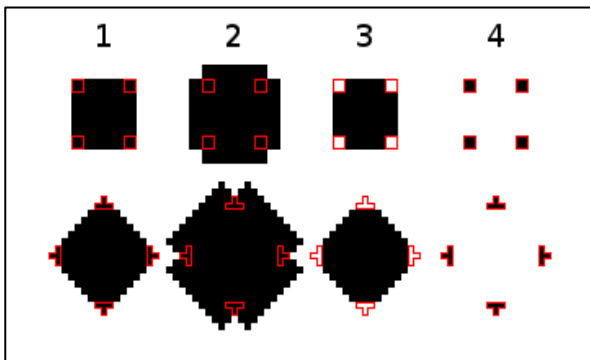


Figure 4. Corner detection steps

Fig. 5 demonstrates the dilate-erode algorithm used on an image segment that contains text. The dilate steps are substantially whiter than their inputs, because the appropriate kernel is used to expand white pixels. Then the erode steps partially reverse this whitening effect by expanding darker pixels. The result is the pixels with the largest differences between the original image and the result image. Fig. 5 shows corners detected on an image segment with text.

Our previous NL localization algorithm [5] was based on the assumption that the NL exists in the image and is horizontally or vertically aligned with the smartphone's camera. Unfortunately, these conditions sometimes do not hold in the real world due to shaking hands or failing eyesight. The exact problem that the new algorithm addresses is twofold. Does a given input image contain a skewed NL? And, if so, within which aligned rectangular area can the NL be localized? In this investigation, a skewed NL is one which has been rotated away from the vertical alignment axis by up to 35 to 40 degrees in either direction, i.e., left or right. An additional objective is to decrease processing time for each frame to about one second.

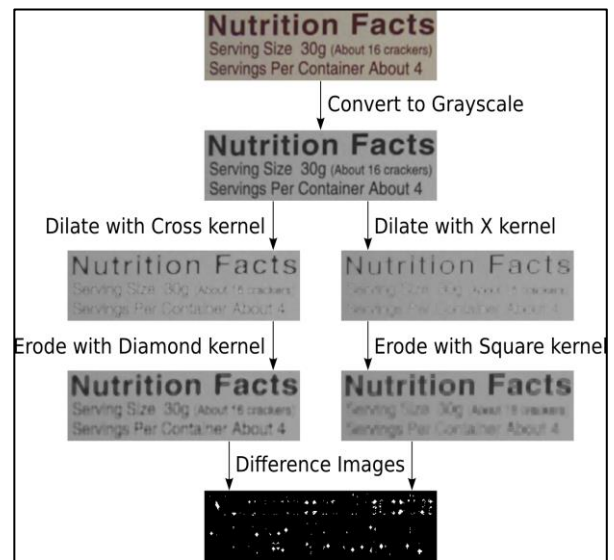


Figure 5. Corner detection for text spotting

B. Corner Detection and Analysis

Before the proper NL localization begins, a rotation correction step is performed to align inputs which may be only nearly aligned. This correction is performed by taking advantage of high numbers of horizontal lines found within NLs. All detected lines that are horizontal within 35 to 40 degrees in either direction (i.e., up or down) are used to compute an average horizontal rotation. This rotation is used to perform the appropriate correcting rotation. Corner detection is executed after the rotation. The dilate-erode corner detector is applied to retrieve a two-dimensional bitmap where *true* white pixels correspond to detected corners and all other *false* pixels are black. Fig. 5 (right) shows the corners detected in the frame shown in Fig. 5 (left).

The dilate-erode corner detector is used specifically because of its high sensitivity to contrasted text, which is why we assume that the region is bounded by these edges contains a large amount of text. Areas of the input image which are not in focus do not produce a large amount of corner detection results and tend not to lie within the needed projection boundaries.

Two projections are computed after the corners are detected. The projections are sums of the true pixels for each row and column. The image row projection has an entry for each row in the image while the image column projection has an entry for each column in the image. The purpose of these projections is to determine boundaries for the top, bottom, left, and right boundaries of the region in which most corners lie. Each value of the projection is averaged together and a projection threshold is set to twice the average. Once a projection threshold is selected, the first and last indexes of each projection greater than the threshold are selected as the boundaries of that projection.

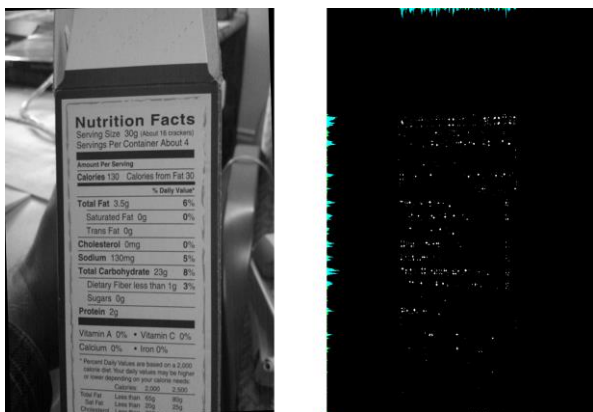


Figure 5. NL (left); detected corners (right)

C. Selection of Boundary Lines

After the four corner projections have been computed, the next step is to select the Hough lines that are closest to the boundaries selected on the basis of the four corner projections. In two images of Fig. 6 the four light blue lines are the lines drawn on the basis of the four corner projection counts. The dark blue lines show the lines detected by the Hough transform. In Fig. 6 (left), the bottom light blue line is initially chosen conservatively where the row corner projections drop below a threshold. If there is evidence that there are some corners present after the initially selected bottom lines, the bottom line is moved as far below as possible, as shown in Fig. 6 (right).



Figure 6. Initial boundaries (left); Final boundaries (right)

When the bounded area is not perfectly rectangular, which makes integration with later analysis where a rectangular area is expected to be less straightforward. To overcome this problem, a rectangle is placed around the selected Hough boundary lines. After the four intersection coordinates are computed, their components are compared and combined to find a smallest rectangle that fits around the bounded area. This rectangle is the final result of the NL localization algorithm. As was stated before, the four corners found by the algorithm can be passed to other algorithms such as row dividing, word splitting, and OCR. Row dividing, world splitting, and OCR are

beyond the scope of this paper. Fig. 7 shows a skewed NL localized by our algorithm.

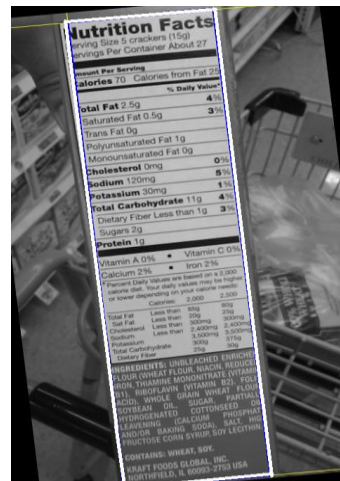


Figure 7. Localized Skewed NL

IV. Experiments & Results

A. Experiment Design

We assembled 378 images captured from a Google Nexus 7 Android 4.3 smartphone during a typical shopping session at a local supermarket. Of these images, 266 contained an NL and 112 did not. Our skewed NL localization algorithm was implemented and tested on the same platform with these images.



Figure 8. Complete (left) and partial (right) true positives

We manually categorized the results into five categories: complete true positives, partial true positives, true negatives, false positives, and false negatives. A complete true positive is an image where a complete NL was localized. A partial true positive is an image where only a part of the NL was localized by the algorithm. Fig. 8 shows examples of complete and partial true positives.

Fig. 9 shows another example of complete and partial true positives. The image on the left was classified as a complete true positive, because the part of the NL that was not detected

is insignificant and will likely be fixed through simple padding in subsequent processing. The image on the right, on the other hand, was classified as a partial true positive. While the localized area does contain most of the NL, some essential text in the left part of the NL is excluded, which will likely cause failure in subsequent processing. In Fig. 10, the left image technically does not include the entire NL, because the list of ingredients is only partially included. However, we classified it as a complete true positive since it includes the entire table on nutrition facts. The right image of Fig. 10, on the other hand, is classified as a partial true positive, because some parts of the nutrition facts table is not included in the detected area.



Figure 9. Complete (left) and partial (right) true positives

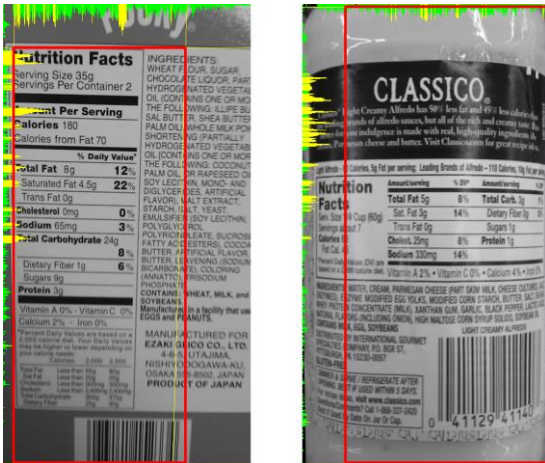


Figure 10. Complete (left) and partial (right) true positives

B. Results

Of the 266 images that contained NLs, 83 were classified as complete true positives and 27 were classified as partial true positives, which gives a total true positive rate of 42% and a false negative rate of 58%. All test images with no NLs were classified as true negatives. The remainder of our analysis was done via precision, recall, and specificity, and accuracy.

Precision is the percentage of complete true positive matches out of all true positive matches. Recall is the percent-

age of true positive matches out of all possible positive matches. Specificity is the percentage of true negative matches out of all possible negative matches. Accuracy is the percentage of true matches out of all possible matches

Table I. NL Localization Results

PR	TR	CR	PR	SP	ACC
0.7632	0.422	0.3580	0.1475	1.0	0.5916

Table I gives the NL localization results where PR stands for “precision,” TR - for “total recall,” CR – for “complete recall,” PR – for “partial recall,” SP – for “specificity,” and ACC – for “accuracy.”. While total and complete recall numbers are somewhat low, this is a necessary trade-off of maximizing specificity. Recall from Section I that we have designed our algorithm to maximize specificity. In other words, the algorithm is less unlikely to detect NLs in images where no NLs are present than in images where they are present. As we argued above, lower recall and precision may not matter much because of the fast rate at which input images are processed on target devices, but there is definitely room for improvement.

C. Limitations

The majority of false negative matches were caused by blurry images. Blurry images are the result of poor camera focus and instability. Both the Canny edge detector and dilate-erode corner detector require rapid and contrasting changes to identify key points and lines of interest. These points and lines are meant to correspond directly with text and NL borders. These useful data cannot be retrieved from blurry images, which results in run-time detection failures. The only recourse to deal with blurry inputs is improved camera focus and stability, both of which are outside the scope of this algorithm, because it is a hardware problem. It is likely to work better in later models of smartphones. The current implementation on the Android platform attempts to force focus at the image center but this ability to request camera focus is not present in older Android versions. Over time, as device cameras improve and more devices run newer versions of Android, this limitation will have less impact on recall but it will never be fixed entirely.

Bottles, bags, cans, and jars (see Fig. 11) have a large showing in the false negative category due to Hough line detection difficulties. One possibility to get around this limitation is a more rigorous line detection step in which a segmented Hough transform is performed and regions which contain connecting detected lines are grouped together. These grouped regions could be used to warp a curved image into a rectangular area for further analysis.

Smaller grocery packages (see Fig. 12) tend to have irregular NLs that place a large amount of information into tiny spaces. NLs with irregular layouts present an extremely difficult problem for analysis. Our algorithm better handles more traditional NL layouts with generally empty surrounding areas. As a better analysis of corner projections and Hough lines is integrated into our algorithm, it will become possible to classify inputs as definitely traditional or more irregular. If

this classification can work reliably, the method could switch to a much slower and generalized localization to produce better results in this situation while still quickly returning results for more common layouts.



Figure 11. NL with curved lines



Figure 12. Irregular NLs

V. Conclusions

We have made several interesting observations during our experiments. The row and column projects have two distinct patterns. The row projection tends to create evenly spaced short spikes for text in each line of text within the NL while the column projection tends to contain one very large spike where the NL begins at the left due to the sudden influx of detected text. We have not performed any in-depth analysis of these patterns. However, the projection data were collected for each processed image. We plan to do further investigations of these patterns, which will likely allow for run-time detection and corresponding correction of inputs of various rotations. For example, the column projections could be used for greater accuracy in determining the left and right bounds of the NL while row projections could be used by later analysis steps such as row division. Certain projection profiles could eventu-

ally be used to select customized localization approaches at run time.

During our experiments with an iterative development of this algorithm, we took note of several possible improvements that could positively affect the algorithm's performance. First, since input images are generally not square, the HT returns more results for lines in the longer dimension, because they are more likely to pass the threshold. Consequently, specifying different thresholds for the two dimensions and combining them for various rotations may produce more consistent results.

Second, since only those Hough lines that are nearly vertical or horizontal are of use to this method, improvements can be made by only allocating bins for those Θ and ρ combinations that are considered important. Fewer bins means less memory to track all of them and fewer tests to determine which bins need to be incremented for a given input.

Third, both row and column corner projections tend to produce distinct patterns which could be used to determine better boundaries. After collecting a large amount of typical projections, further analysis can be performed to find generalizations resulting in a faster method to improve boundary selection.

Fourth, in principle, a much more intensive HT method can be developed that would divide the image into a grid of smaller segments and perform a separate HT within each segment. One advantage of this approach is to look for the skewed, curved, or even zigzagging lines between segments that could actually be connected into a longer line. While the performance penalty of this method could be quite high, it could allow for the detection and de-warping of oddly shaped NLs. Finally, a more careful analysis of the found Hough lines during the early rotation correction could allow us to detect and localize NLs of all possible rotations, not just skewed ones.

The U.S. Food and Drug Administration recently proposed some changes to the design of NLs on product packages [14]. The new design is expected to change how serving sizes are calculated and displayed. Percent daily values are expected to shift to the left side of the NL, which allegedly will make them easier to read. The new design will also require information about added sugars as well as the counts for Vitamin D and potassium. We would like to emphasize that this redesign, which is expected to take at least two years, will not impact the proposed algorithm, because the main tabular components of the new NL design will remain the same. The nutritional information in the new NLs will still be presented textually in rows and columns. Therefore, the corner and line detection and their projections will work as they work on the current NL design.

References

- [1] Anding, R. *Nutrition Made Clear*. The Great Courses, Chantilly, VA, 2009.
- [2] Rubin, A. L. *Diabetes for Dummies*. 3rd Edition, Wiley, Publishing, Inc. Hoboken, New Jersey, 2008.
- [3] Nutrition Labeling and Education Act of 1990. http://en.wikipedia.org/wiki/Nutrition_Labeling_and_Education_Act_of_1990.

- [4] Food Labelling to Advance Better Education for Life. Avail. at www.flabel.org/en.
- [5] Kulyukin, V., Kutiyawala, A., Zaman, T., and Clyde, S. "Vision-based localization and text chunking of nutrition fact tables on android smartphones," In Proc. International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICCV 2013), pp. 314-320, ISBN 1-60132-252-6, CSREA Press, Las Vegas, NV, USA, 2013.
- [6] Nicholson, J. and Kulyukin, V. "ShopTalk: Independent Blind Shopping = Verbal Route Directions + Barcode Scans." In Proceedings of the 30-th Annual Conference of the Rehabilitation Engineering and Assistive Technology Society of North America (RESNA 2007), June 2007, Phoenix, Arizona. Avail. on CD-ROM.
- [7] Kulyukin, V. and Kutiyawala, A. "Accessible Shopping Systems for Blind and Visually Impaired Individuals: Design Requirements and the State of the Art." The Open Rehabilitation Journal, ISSN: 1874-9437, Volume 2, 2010, pp. 158-168, DOI: 10.2174/1874943701003010158.
- [8] Kulyukin, V., Kutiyawala, A., and Zaman, T. "Eyes-Free Barcode Detection on Smartphones with Niblack's Binarization and Support Vector Machines." In Proceedings of the 16-th International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICCV 2012), Vol. I, pp. 284-290, CSREA Press, July 16-19, 2012, Las Vegas, Nevada, USA. ISBN: 1-60132-223-2, 1-60132-224-0.
- [9] Kulyukin, V. and Zaman T. "Vision-Based Localization of Skewed UPC Barcodes on Smartphones." In Proceedings of the International Conference on Image Processing, Computer Vision, & Pattern Recognition (ICCV 2013), pp. 344-350, pp. 314-320, ISBN 1-60132-252-6, CSREA Press, Las Vegas, NV, USA.
- [10] B. J. Fog. "A behavior model for persuasive design," In Proc. 4th International Conference on Persuasive Technology, Article 40, ACM, New York, USA, 2009.
- [11] Canny, J.F. "A Computational approach to edge detection." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, 1986, pp. 679-698.
- [12] Duda, R. O. and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," Comm. ACM, Vol. 15, pp. 11-15, January, 1972.
- [13] Laganiere, R. OpenCV 2 Computer Vision Application Programming Cookbook. Packt Publishing Ltd, 2011.
- [14] S. Tavernise. "New F.D.A nutrition labels would make 'serving sizes' reflect actual servings." New York Times, Feb. 27, 2014.

Automated Hair Color Determination

Daniel S. Rosen¹, and Cambron Carter²

¹Director, Imaging Technology, GumGum Inc., Santa Monica, CA, US

²Image Scientist, GumGum Inc., Santa Monica, CA, US

Abstract— *The detection of human features utilizing computer vision techniques can provide significant information for exploitation of image content. Identification of human hair and its color is known to be of use for a variety of endeavors including targeting advertisements of hair care products. The daily volume of imagery which must be processed for advertising as well as the uncontrolled environment in which they are typically captured, negates the use of semi-automated techniques. A method of automated hair color determination which achieves high accuracy is presented.*

Keywords- segmentation; parameter estimation; expectation maximization; hair detection; heuristics

1 Introduction

There are two main pigments found in human hair (both being melanins): eumelanin, which gives color to brown or black hair and pheomelanin, which produces the color in blonde or red hair. Thus, natural hair color will only occur as shades/combinations of blonde, red, brown and black, with shades of gray (all the way to white) occurring as melanin production slows and/or stops. While hair color may also be influenced by the optical effects of light reflecting off the surfaces of the different hair layers, natural hair will never contain blue or green. However, since white hair is devoid of pigment, white or gray hair may appear to have a bluish color due to the refraction of light.

The GumGum hair color determination algorithm (GHCD) was created for the purpose of intelligently targeting hair care products in advertisements. As such, “ground truth” is defined as what the average observer would consider the hair color to be, not necessarily what the person in the image states about their hair color. In addition, “unnatural” hair colors, such as blue, purple and green are classified as unknown.

Hair detection has been performed based on analysis of a hair mask obtained by segmentation [1], and by segmentation using frequential and color analysis followed by matting [2]. While producing excellent results, the former technique lacks robustness—suffering accuracy when the hair and background are non-uniform. The primary disadvantage of the latter technique is that it requires a manual seeding step. A technique has been developed utilizing Markov random fields [3], but this also requires seeding and further

relies on careful training of a location prior model. Finally, a technique has been developed using active shapes based on training a hair shape model [4]. While yielding good results, this technique is susceptible to error in cases of large shape variation, which may be caused by lighting effects or image geometry.

2 Approach

The GHCD algorithm logic is shown in Figure 1.

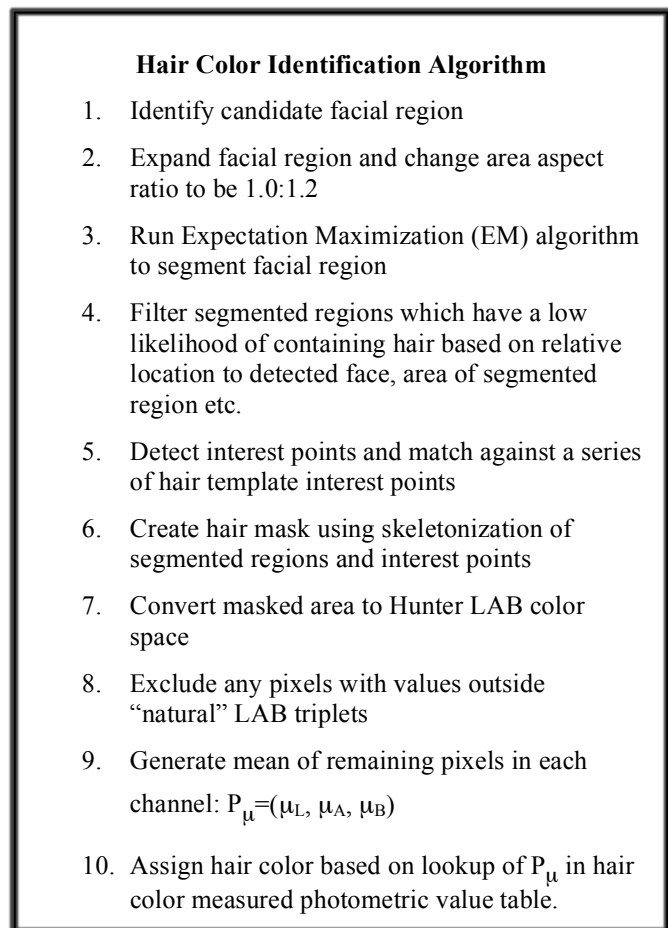


Figure 1.
GHCD pipeline

First, a facial classifier such as that defined by Viola and Jones [5] is utilized to detect all faces within the image. In particular, a frontal face classifier [6] and a profile face classifier [7] are used to detect face candidates. It is understood that other techniques for locating a face, such as eye detection, etc. may be substituted.

Once the face has been located, an expanded area R , (Figure 2) is created, centered at the center of the face detection rectangle (white square in Figure 2) (X_c, Y_c) . This increases the probability of including significant amounts of hair, given that human hair is attached to the human head. With the Viola/Jones detection rectangle as a baseline, we create the expanded search area R with height H and width W given by

$$H = H_v * 2.0 \tag{1}$$

where H_v =height of Viola/Jones rectangle

$$W = W_v * 1.67 \tag{2}$$

where W_v =width of Viola/Jones rectangle



Figure 2.

Input image with the detected face outlined.

The expanded search area R is then sent to an Expectation Maximization (EM) algorithm for segmentation. The EM algorithm is an iterative procedure for finding maximum likelihood estimates of parameters describing statistical processes in cases where the process depends on hidden, random variables, i.e. missing/sparse data. Assuming a data point, Z , we wish to ascertain the likelihood $Z \in z$, where z is representative of a class present within the given data. The EM algorithm iteratively alternates between an expectation step and a maximization step. The expectation step finds the expectation of the log-likelihood current parameter estimates while the maximization step maximizes the expected log-likelihood produced in the expectation step. This process leapfrogs back and forth until converging to stable parameter estimates, which describe the statistical process.

This procedure is straightforward provided $p(X, Z/\theta)$ takes on a closed form. For our purposes, the form of p is assumed to be a Gaussian Mixture Model (GMM), with θ

representing mean, μ , covariance, Σ , and mixing coefficient, α . Given the GMM form, each Expectation step determines the posterior probability based on the current μ , Σ , and α and serves to satisfy the probability that $Z=z$ best represents the hidden (or missing data) given what we have observed. Each Maximization step updates μ , Σ , and α using update equations. As to be expected, the algorithm requires an initialization of θ (μ , Σ , and α) and will increase the log-likelihood of the data that has been observed until a maximum is reached. The EM algorithm logic is shown in Figure 3.

EM Algorithm for Gaussian Mixture Model

1. Initialize parameter vector $\theta^0 = \{\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}$.
2. Begin EM: $t = 0$
3. While the log-likelihood of observed data $\sum_{i=1}^n \log p(x_i | \theta^t)$ is increasing:
 - a. E-step:

$$r_{ij} = p(z_i = j | x_i, \theta) = \frac{\alpha_j p(x_i, \mu_j, \Sigma_j)}{\sum_{j=1}^k \alpha_j p(x_i, \mu_j, \Sigma_j)}$$
 - b. M-step: re-estimate parameters using the update equations:

$$\hat{\alpha}_l = \frac{\sum_{i=1}^n r_{il}}{\sum_{j=1}^k \sum_{i=1}^n r_{ij}} \quad \hat{\mu}_l = \frac{\sum_{i=1}^n r_{il} x_i}{\sum_{i=1}^n r_{il}}$$

$$\hat{\Sigma}_l = \frac{\sum_{i=1}^n r_{il} (x_i - \mu_l)(x_i - \mu_l)^T}{\sum_{i=1}^n r_{il}}$$
 - c. Determine θ^{t+1} based on the re-estimation for each Gaussian in the mixture model.
4. $\theta^t = \theta^{t+1}$; $t = t + 1$; jump to a.

Figure 3.

Summary of the EM Algorithm

The EM algorithm is slightly sensitive to initialization as it may converge to a local maximum rather than the true, global maximum of the log-likelihood of the observed data. Figure 4 illustrates iteratively fitting GMMs with 1, 2, 3, 4 and 5 Gaussians profiles respectively to a raw, 1-channel histogram (representative of our observed data).

By iteratively employing the EM algorithm, an image may be adequately partitioned into regions of interest, which

can be treated as a reliable, initial blind segmentation. EM generates a set of candidate memberships based on the number of Gaussians fitted to the parameter space. Experimentation has led to the choice of a GMM utilizing up to 7 Gaussians. Given subsequent segmentations from this procedure, we now wish to logically obtain those segmentations which best represent regions containing hair.

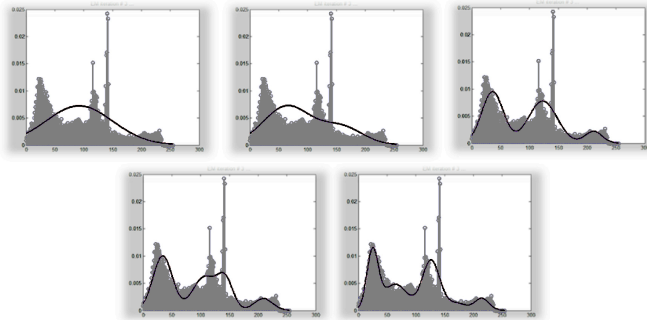


Figure 4.
Iteratively fitting 1,2,3,4, and 5 Gaussians to some observed data using EM.

Accepting a segmented region as a possible hair candidate begins with narrow-banding around an approximated ellipse, which is assumed to outline the facial region. This follows as a direct result of face detection. Using this narrow-band in combination with a connected-component labeling of the segmented region, candidates are taken as those connected components, which intersect the narrow-band. These candidates must then be further processed to determine the likelihood that they, in fact, represent hair. To address this, the pipeline in Figure 5 is followed for each candidate hair region:

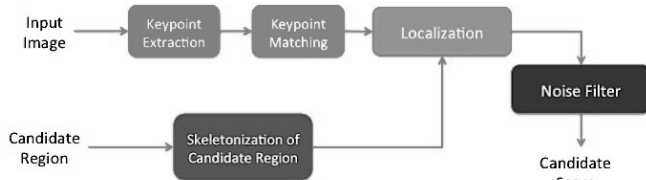


Figure 5.
Pipeline to filter segmented regions deemed to be “non-hair” regions.

Each input image undergoes feature extraction, where those features may come from any state-of-the-art method including HoG, SURF, SIFT, Daisy, etc. These features are then matched with a set of descriptors generated offline using a database of known hair templates. These templates were chosen to exploit the extreme variation in the appearance of human hair, as well as the extreme variations in imaging

conditions to which images containing human hair appear. A subset of the template database is shown in Figure 6:

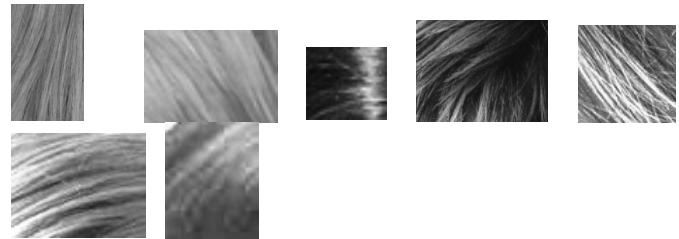


Figure 6.
Subset of hair templates database used to generate feature descriptors offline for the purpose of feature matching with an input image.

In parallel, each candidate region obtained from the EM algorithm is skeletonized using basic morphological operations. Merging information from the feature matching and skeletonization results in a score. This score represents the number of feature points at or within a rigid threshold distance to the nearest index along the candidate skeleton. This process is illustrated visually below:

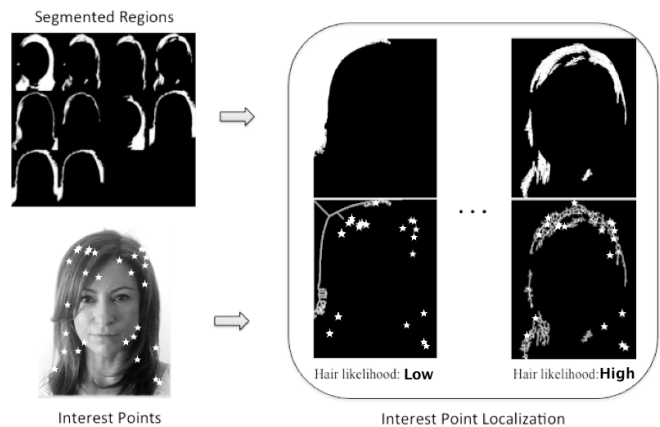


Figure 7.
Top Left : Segmented regions obtained from EM.
Bottom Left : Interest point detection.
Right : Localization of interest points based on average distance to skeletonized region.

The process of localizing interest points acts as a scoring system for each segmented region. Those skeletons with a large average distance to the interest points will be considered non-hair regions. This technique, combined with heuristic selections based on human physiology, yields a final hair candidate mask as in Figure 8.

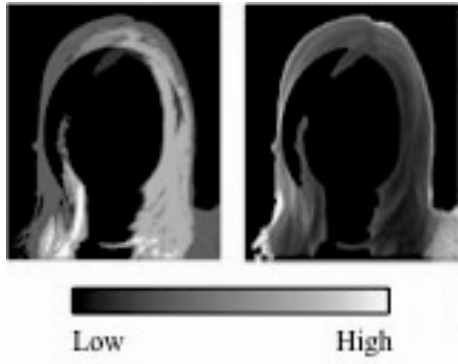


Figure 8.

Left : Original hair mask (brighter equals higher hair likelihood) Right : Masking the raw input image.

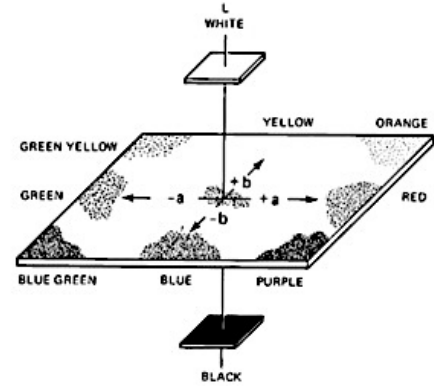


Figure 9.

Hunter Lab color space.

In order to negate the effects of illumination, the CIELab color space was originally chosen for the final hair color detection and identification. However, it was found that the cosmetic industry, as well as the paint and pigment industries, performs all their research and QA colorimetry in the Hunter Lab (HLab) space. So, it was decided to convert RGB to HLab instead of CIELab for all further processing.

It can be seen from Figure 9, that the range of values for natural hair color is constrained to positive and slightly negative values of a and b , e.g. $-10.0 \leq a$ and $-10.0 \leq b$.

The Hunter Lab space, shown in Figure 9, was developed in 1948 by R. S. Hunter as a uniform color space which could be read directly from a photoelectric colorimeter (tristimulus method). Values in this space are defined by the following formulas:

These constraints on a and b are used to detect and tag pixels P in the masked image which are deemed invalid by the simple test:

$$L = 100 \cdot \sqrt{\frac{Y}{Y_0}} \quad (3)$$

$$a = 175 \cdot \left[\sqrt{\frac{0.0102X_0}{(Y/Y_0)}} \right] \cdot \left(\frac{X}{X_0} - \frac{Y}{Y_0} \right) \quad (4)$$

$$b = 70 \cdot \left[\sqrt{\frac{0.00847Z_0}{(Y/Y_0)}} \right] \cdot \left(\frac{Y}{Y_0} - \frac{Z}{Z_0} \right) \quad (5)$$

$$\text{Given } P=\{L,a,b\}, P \text{ is valid} \Leftrightarrow -10.0 \leq a \wedge -10.0 \leq b \quad (T1)$$

The results of applying (T1) to the masked image are shown in Figure 10, with the convex hull of candidate pixels also being shown. Once the candidate region has been formed, the mean values of L, a, b ($\hat{L}, \hat{a}, \hat{b}$) are determined by

$$\hat{L} = \frac{(\sum_j \sum_k L_{j,k})}{n} \quad (9)$$

$$\hat{a} = \frac{(\sum_j \sum_k a_{j,k})}{n} \quad (10)$$

$$\hat{b} = \frac{(\sum_j \sum_k b_{j,k})}{n} \quad (11)$$

where: X, Y, Z = Tristimulus values of the specimen.
 X_0, Y_0, Z_0 = Tristimulus values of the perfect reflecting diffuser

For the 2° Standard Observer and Standard Illuminant C, the above equations would become:

$$L = 100 \cdot \sqrt{Y} \quad (6)$$

$$a = \frac{17.5 \cdot (1.02X - Y)}{\sqrt{Y}} \quad (7)$$

$$b = \frac{70 \cdot (Y - 0.847Z)}{\sqrt{Y}} \quad (8)$$



Figure 10.

Application of hair mask to input image. Black pixels represent non-hair and white contour represents the convex hull of hair region.

The hair care industry has devoted many decades of research towards developing their products. In the course of these efforts, many controlled measurements of natural and dyed hair color have been made. As previously noted, these measurements are performed in the Hunter Lab color space.

A detailed table of color measurements was obtained [8] and consolidated slightly to reduce the number of possible hair color assignments from 68 to 62. A small portion of the hair color table is shown in Figure 9.

Color	L min	L max	a min	a max	b min	b max
Black	0.0	14.0	-10.0	3.0	-10.0	5.0
Very Dark Brown – cool overtones	14.0	16.0	-10.0	3.0	-10.0	1.0
Very Dark Brown	14.0	16.0	-10.0	3.0	1.0	1.25
Very Dark Brown – warm overtones	14.0	16.0	-10.0	3.0	1.25	3.0
Dark Auburn – cool overtones	16.0	19.0	2.0	3.0	-10.0	2.7
Lightest Blonde	40.0	50.0	1.8	5.0	9.0	10.0
Red Blonde	27.0	40.0	7.0	30.0	6.0	30.0
Red	19.0	22.0	2.0	30.0	3.5	4.0

Figure 11.

Hair color table used for identifying hair color. Values are represented in HLab.

There is significant variation in the actual description of various shades of hair color. For example, “chestnut” is usually meant as a “brownish auburn”, with no objective definition of “brownish”. Because of this, GHCD identifies hair color as a triplet (C,S,O) where C = color, S = shading notation, and O = overtone. These are defined as:

$(C \in \mathbf{n}) : \mathbf{n} = \{\text{black, brown, blonde, red, auburn, gray}\}$

$(S \in \mathbf{l}) : \mathbf{l} = \{\text{darkest, very dark, dark, medium, lightest, very light, light}\}$

$(O \in \mathbf{t}) : \mathbf{t} = \{\text{cool, medium, warm}\}$

The triplet $(\hat{L}, \hat{a}, \hat{b})$ is compared to the measurement table and the hair triplet (C,S,O) is thus determined. Figure 12 shows a final hair color determination.



Figure 12.

Final hair color determination.

3 Results and Further Work

A set of 675, random images (containing faces) were chosen from the Internet and hair color was identified in each by GHCD. 602 hair color identifications by GHCD agreed with manual identifications, an 89.2% accuracy. 13 images of the 602 GHCD classified were disputed; e.g. hair color identification could not be agreed upon by manual observers. GHCD incorrectly identified the hair color in 73 images versus the manual identification, a 10.8% error rate.

GHCD has been shown to be an effective algorithm for the automated determination of hair color. Our research has shown that there are three significant sources of error. First, is the segmentation of hair versus non-hair regions, second is the effect of extreme illumination variations in the image and third, the natural or artificial variations which can occur in a person's hair, e.g. “highlights” added to hair.

Moving forward, work still remains to improve the hair segmentation procedure and we are investigating techniques, which can detect and identify significant variations of hair color in a single individual.

4 References

- [1] Y. Yacoob and L.S. Davis, "Detection and analysis of hair", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1164-1169, 2006
- [2] Rousset, C. and Colon, P.Y. "Frequential and Color Analysis for Hair Mask Segmentation", 2008
- [3] Lee, Kuang-chih and Anguleov, Dragomir and Sumengen, Baris and Gokturk, Salih, "Markov Random Field Models for Hair and Face Segmentation", *Proceeding of 8th IEEE International Conference on Automatic Face & Gesture Recogniton*, 2008
- [4] Julian, P. and Dehais, C. and Lauze, F. and Charvillat, V. and Bartoli, A. and Choukroun, A., "Automatic Hair Detection in the Wild", *ICPR 2010 - 20th International Conference on Pattern Recognition*, 4617-4620, 2010
- [5] Viola, P., & Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. *IEEE Conference CVPR. 1*, pp. 511-518. IEEE.
- [6] Lienhart, R. (2003). 20x20 gentle adaboost frontal face detector. *OpenCV*. Intel.
- [7] Bradley, D. (2003). 20x20 profile face detector. *OpenCV*. Princeton University / Intel.
- [8] MacFarlane, Darby and MacFarlane, David and Billmeyer, Fred, "Method and Apparatus for Hair Color Characterization and Treatment", WO 96/41139

Automated Area Defect Inspection of Touch Panels Using Computer Vision

Hong-Dar Lin, Jen-Miao Li

Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung, 41349, Taiwan

Abstract – Touch panels (TP) are widely used in various electronic products. It is a difficult inspection task when defects embedded in surfaces of TPs with structural textures. A common surface defect type called area defects includes dirt, water marks, bubbles, and other defects with larger size. Such defects have low contrast, brightness with gradual changes, irregular and non-directional shapes, and there may be both bright and dark flaws co-existing in a region. Therefore, this study proposes an automated detection method to inspect the area defects on touch panels. The proposed method applies the Haar Wavelet transform with flat zone filtering operation to remove the structural textures of background through filtering an approximated sub-image of a decomposed image in wavelet domain. Then, the filtered image is transformed back to spatial domain. Finally, the restored image can be easily segmented to into three categories namely dark defects, bright defects, and background by using a simple statistical histogram method. Experimental results show that the proposed method achieves a high 91.82% defect detection rate, a low 4.36% false alarm rate.

Keywords: Industrial inspection; touch panels; area defects; computer vision system; Wavelet transform, flat zone filter.

1 Introduction

With the development of smart-phones, general touchtone phone gradually being replaced in order to stimulate a wave of touch screen. Touch panel not only for mobile phones and even extends to the computer, television, camera, handheld game consoles and other 3C products, the increasing demand of touch panels promotes the development of touch panel industry. Currently touch technology of touch panels can be primarily divided into resistive, capacitive, optical, electromagnetic, ultrasonic types, which the resistive type is an earlier technology. Capacitive Touch Panels (CTPs) have advantages of water-proof, stain-proof, scratch-proof, fast response, anti-UV, etc. Therefore, the CTP products have an unshakable market position in touch panels.

The surface defects are usually classified into linear and area types of defects [1]. The area type includes dirt, water marks, bubbles and other defects with larger size. Such defects have low contrast, brightness with gradual changes,

irregular and non-directional shapes, and there may be both bright and dark flaws co-existing in a region. This kind of defects compared with the linear type is more complicated to identify its regularity. Therefore, this study proposes an automated detection method to inspect the area defects on touch panels. Figure 1 shows the normal and defective images of CTP surfaces with directional textures. The directional textures reveal lattice shapes with lines in four directions (horizontal, vertical, and two diagonals). These background textures make the defect inspection tasks more difficult when area defects embedded in the surfaces of directional textures.

Currently, difficulties exist in precisely inspecting area defects by machine vision systems because when product images are being captured, the region of an area defect could expand, shrink or even disappear due to uneven illumination of the environment, complex texture of the product surface, and so on. Therefore, we propose a wavelet transform based image restoration approach to overcome the difficulties of automated touch panel area defect inspection.

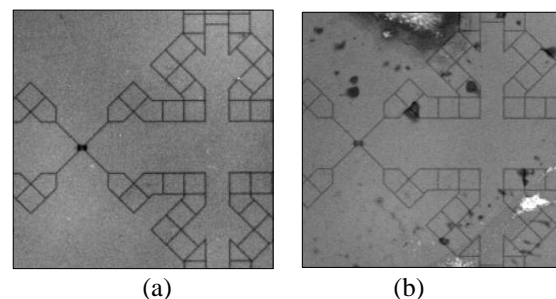


Figure 1. The CTP images with directional texture: (a) a normal image; (b) a defective image with area defects.

2 Automated defect inspections

Automated inspection of surface flaws has become a critical task for manufacturers who strive to improve product quality and production efficiency [2, 3]. Flaw detection techniques, generally classified into the spatial domain and the frequency domain, compute a set of textural features in a sliding window and search for significant local deviations among the feature values. Latif-Amet *et al.* [4] presented wavelet theory and co-occurrence matrices for detection of defects encountered in textile images and classify each sub-

window as defective or non-defective with a Mahalanobis distance. Cho *et al.* [5] applied the adaptive threshold technique and morphology method to detect defects from images of uniform fabrics for developing a real-time vision system. Perng *et al.* [6] used three extracted features of segmented fringe images and a control chart procedure to inspect optical defects in quasi-contact lenses.

Automated thresholding has also been widely used in the computer vision applications for automated optical inspection of defects [7]. The Otsu method [8] is one of the better threshold selection methods for general real world images with respect to uniformity and shape measures. This method selects threshold values that maximize the between-class variances of the histogram. The Otsu method is optimal for thresholding large objects from the background. It provides satisfactory results for thresholding an image with a histogram of bimodal distribution. Ng [9] revised the Otsu method for selecting optimal threshold values for both unimodal and bimodal distributions, and tested the performance of the revised method on common defect detection applications.

Fourier transform, wavelet transform and Gabor transform are common texture analysis techniques used in the frequency domain [10]. Chan and Pang [11] used the Fourier transform to detect fabric defects. Tsai and Hsiao [12] proposed a wavelet transform based approach for inspecting local defects embedded in homogeneous textured surfaces. Lin [13] developed a wavelet-based multivariate statistical approach to automatically inspect ripple defects with low intensity contrast in the surface barrier layer chips of ceramic capacitors. Kumar and Pang [14] proposed supervised and unsupervised defect detection approaches for automated inspection of textile fabrics using Gabor wavelet features. Lin and Jiang [15] combined discrete cosine transform and grey relational analysis technique to inspect surface defects on encapsulations of light emitting diodes. Also, Lin [16] further developed a novel approach that applies discrete cosine transform decomposition and cumulative sum techniques for the detection of tiny defects on passive component chips.

Tsai and Hsieh [17] proposed a global image restoration scheme using the Fourier transform and Hough transform for the automatic inspection of defects in directionally textured surfaces. Perng and Chen [18] developed a nonnegative matrix factorization based approach for automatically inspecting the defects in directional texture surfaces. As to inspecting defects of touch panels, Chen *et al.* [19] introduced an automated optical inspection system for analogical RTP. This system integrates mechanism, electrical control and machine vision, and applies digital image processing method to inspect defect of the RTP. The RTP has the texture of periodic spacers in spatial domain image and results in periodic dots in Fourier spectrum.

In this research, we explore the area defect inspection of the popular CTP products. It is difficult to precisely detect area defects embedded in the complicated directional textures. Therefore, we present a global image restoration scheme using the wavelet transform and flat zone filtering process for area defect detection on CTP images. This scheme does not need feature extraction and template matching processes.

3 Proposed method

This research proposes a Wavelet Transform (WT) based flat zone filtering approach to inspect area defects of touch panels. When a touch panel image with four different directional line patterns of background texture is transformed to wavelet domain, the directional textures of background can be removed through filtering the approximated sub-image of the next decomposition level of Wavelet transform. Then, the filtered image is transformed back to spatial domain. Finally, the restored image with enhanced defects can be easily segmented into three categories, dark defects, bright defects, and background, by using a simple statistical histogram method and some features of the detected defects are extracted.

3.1 Wavelet transform

Wavelet transform provides a convenient way to obtain a multi-resolution representation, from which texture features can be easily extracted. We use the Haar wavelet transform to conduct image transformation for extraction of image features, because the merits of Haar wavelet transform include local image processing, simple calculations, high speed processing, memory efficiency, and multiple image information [20]-[23]. The Haar wavelet transform is one of the simplest and basic wavelet transformations. The Haar transform can be computed stepwise by the mean value and half of the differences of the tristimulus values of two contiguous pixels. Based on the transfer concept of the 1-D space, the Haar wavelet transform can process a 2-D image of $(M \times N)$ pixels in the following way:

Row transfer:

$$\begin{cases} f_R(i, j) = \left[\frac{f(i, 2j) + f(i, 2j+1)}{2} \right], \\ f_R(i, j + \left[\frac{N}{2} \right]) = \left[\frac{f(i, 2j) - f(i, 2j+1)}{2} \right], \\ \text{where } 0 \leq i \leq (M-1), 0 \leq j \leq \left[\frac{N}{2} \right] - 1, [] \text{ is Gauss symbol.} \end{cases}$$

Column transfer:

$$\begin{cases} f_C(i, j) = \left[\frac{f_R(2i, j) + f_R(2i+1, j)}{2} \right], \\ f_C(i + \left[\frac{M}{2} \right], j) = \left[\frac{f_R(2i, j) - f_R(2i+1, j)}{2} \right], \\ \text{where } 0 \leq i \leq \left[\frac{M}{2} \right] - 1, 0 \leq j \leq (N-1). \end{cases} \quad (1)$$

In the above expressions (Eq. (1)), $f(i, j)$ represents an original image, $f_r(i, j)$ the row transfer function of $f(i, j)$, and $f_c(i, j)$ the column transfer function of $f_r(i, j)$. As $f_c(i, j)$ is also the outcome of the wavelet decomposition of $f(i, j)$, the outcomes of a wavelet transform can be defined as:

$$\begin{cases} A(i, j) = f_c(i, j); & D_1(i, j) = f_c(i, j + \lfloor \frac{N}{2} \rfloor); \\ D_2(i, j) = f_c(i + \lfloor \frac{M}{2} \rfloor, j); & D_3(i, j) = f_c(i + \lfloor \frac{M}{2} \rfloor, j + \lfloor \frac{N}{2} \rfloor); \\ \text{where } 0 \leq i \leq \lfloor \frac{M}{2} \rfloor - 1, 0 \leq j \leq \lfloor \frac{N}{2} \rfloor - 1. \end{cases} \quad (2)$$

One level of wavelet decomposition generates one approximated sub-image and three detailed sub-images that contain fine structures with horizontal, vertical, and diagonal orientations. These four sub-images, each of which has a size of $(M/2 \times N/2)$ pixels, form the wavelet characteristics. The proposed method extracts the four textural features of one-level wavelet decomposition to detect water-drop blemishes. Multi-level wavelet decomposition generates coarser representation of the original image. A large number of decomposition levels will result in the fusion effect for the blemishes and may cause localization error of the detected defect [24]. Figure 2 shows the wavelet transforms among three decomposition levels of a structural texture pattern with area defects. Figure 2(a), 2(b), and 2(c) are the transformed images with the first, second, and third decomposition levels, respectively.

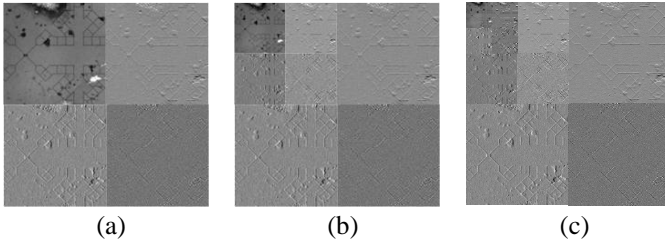


Figure 2. Decomposition of Fig. 1(a) using Haar wavelet at three decomposition levels: (a) first decomposition level; (b) second decomposition level; (c) third decomposition level.

3.2 Flat zone filtering

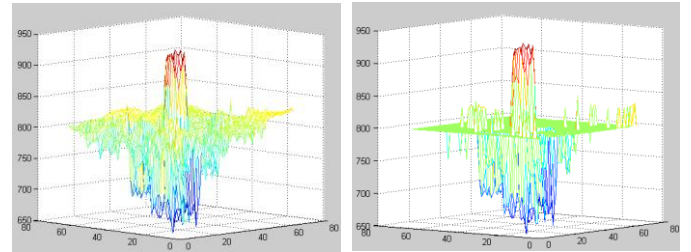
The multi-resolution wavelet technique transforms images into a representation where both spatial and frequency information existing. It is suitable for describing local changes in a homogeneous textured image. For one level of wavelet decomposition, we obtain one approximated sub-image and three detailed sub-images which contain fine structures with horizontal, vertical and diagonal orientations. By properly filtering the approximated sub-image in different decomposition levels for backward wavelet transform, the reconstructed image will remove regular, repetitive texture patterns and enhance only local defects. A simple

thresholding can then be used to discriminate between defective regions and homogeneous regions in the reconstructed image.

Comparing with regular band filtering, the proposed flat zone filtering uses the mean value of overall frequency instead of zero to replace the frequency components within the selected zone area in wavelet frequency domain. The flat zone filtering approach is to reduce the variation between background and texture. The main purpose is to remove the background (directional textures) which with small changes in grayscales, and reserve the area defects which with larger changes in grayscales. We use the control limit concept of \bar{x} control chart to determine the flat zone area $\bar{x} \pm k\sigma$ and replace all of the frequency components within the region by the frequency mean value \bar{x} . This can be written as,

$$\begin{cases} \text{image}(i, j) = \bar{x}, & \bar{x} - k\sigma < \text{image}(i, j) < \bar{x} + k\sigma \\ \text{image}(i, j) = \text{image}(i, j), & \text{otherwise} \end{cases} \quad (3)$$

Figure 3 shows 3-D WT spectrum diagrams which are the before and after plots of the flat zone filtering are conducted. Figure 3(a) is a 3-D diagram of a WT spectrum and Figure 3(b) is a 3-D diagram of a WT spectrum with flat zone filtering. After the flat zone filtering, the frequency components with large frequency are enhanced in WT domain.



(a) Before flat zone filtering (b) After flat zone filtering

Figure 3. A flat zone filter added into a 3-D WT spectrum diagram: (a) A 3-D diagram of a WT spectrum; (b) A 3-D diagram of a WT spectrum with flat zone filtering.

3.3 Reverse Wavelet transform

After the proper zone is determined, the frequency filtering operation can accurately specify the non-defect low frequency regions and set the values of their frequencies at mean value in the WT domain. Then, we can transform the filtered frequency image back to the spatial domain for further defect separation. In this study, we would like to eliminate all regular, repetitive textures in the reconstructed image by selecting a proper band in approximated sub-images for mean value filtering. Since a structural texture may present high directionality, reconstructing the detailed sub-images with direction emphasis different from that of the regular texture will remove all repetitive, directional patterns in the original image, and preserve only local defects in the

restored image. The repetitive, directional pattern will result in an approximately uniform gray level, whereas the local defects will yield distinct gray levels in the restored image.

3.4 Defect separation

The restored image has uniform gray levels for pixels belonging to homogeneous line regions, but it also gives significantly different gray levels for pixels belonging to inhomogeneous defect areas. The intensity variation in homogeneous regions will be very small, whereas the gray-level variation in inhomogeneous areas will be large with respect to the entire restored image. Therefore, we can use a simple statistical principle to set up the control limits for distinguishing defects from periodic line patterns in the restored image. The restored image $f'(x, y)$ will be approximately a uniform gray-level image if a non-defect surface image is tested. The upper and lower control limits (T_L, T_U) for intensity variation in the restored image are given by

$$T = \mu_{f'} \pm k\sigma_{f'} \quad (4)$$

where T is a threshold for segmenting defects from background, k is a control parameter, $\mu_{f'}$ and $\sigma_{f'}$ are the mean and standard deviation of the testing restored image of size $M \times M$. The resulting three level image $B(x, y)$ for defect separation is

$$B(x, y) = \begin{cases} 127, & f'(x, y) > T_U \\ 255, & T_L \leq f'(x, y) \leq T_U \\ 0, & f'(x, y) < T_L \end{cases} \quad (5)$$

If a gray level value falls within the control limits (T_L, T_U) (in control) then intensity is set to 255 (white) as a background. Otherwise, intensity is set to 0 (black) as a part of dark defect if a gray level is less than the lower control limit (out of control) and intensity is set to 127 (gray) as a part of bright defect if a gray level is more than the upper control limit (out of control).

If a pixel with the gray level falls within the control limits, the pixel is classified as a homogeneous element. Otherwise, it is classified as a defective element. As the defect size to be inspected are generally very small with respect to the entire surface image, $\mu_{f'}$ and $\sigma_{f'}$ can be computed directly from the restored image of a testing image to accommodate the variation of lighting in the inspection environment. All experimental samples demonstrated in this study are based on the $\mu_{f'}$ and $\sigma_{f'}$ from testing images, and the control constant k is set at different values.

The control limits are used to distinguish between homogeneous line patterns and defects in a filtered restored image. The upper and lower control limits of gray levels in a restored image are placed at a distance $k\sigma_{f'}$ from the mean

$\mu_{f'}$. Figure 4 depicts the resulting three level images without and with WT based flat zone filtering of the restored images in Fig. 4(a) and Fig. 4(b), where pixels with gray levels falling outside the control limits are represented by black and gray intensities (defective regions), and the ones falling within the limits are represented by white intensity (homogeneous regions). There are many false alarms existing in Fig. 4(a) and most of the area defects are correctly detected in Fig. 4(b). This indicates the proposed WT based flat zone filtering method has the ability to precisely locate the area defects in directional textures. In addition, selecting a proper control parameter results in correctly discriminating defects from normal regions but an improper control parameter produces many erroneously detecting normal regions as defects. A smaller constant value k gives a tight control and may result in false alarms. A larger constant value k gives a loose control and may generate missing alarms.

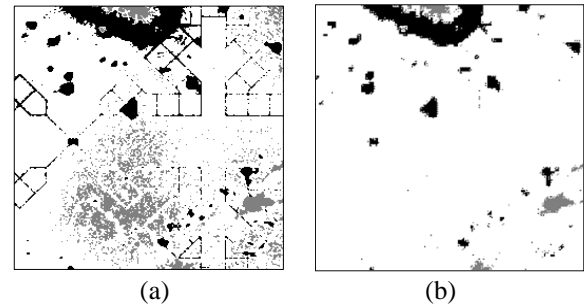


Figure 4. The resulting three level images of defect detection with and without WT based flat zone filtering: (a) without WT filtering; (b) with WT filtering.

4 Experiments and analyses

To evaluate performance of the proposed approach, experiments were conducted on real capacitive touch panels, provided by a touch panel manufacturing company. The CTP images (70) with thickness 0.78mm, of which 32 have no defects and 38 have various area defects, were tested. All samples were randomly selected from manufacturing process of touch panels. Each image of the surface has a size of 256 x 256 pixels and a gray level of 8 bits. The area defect detection algorithm is edited and executed on the R2009b version of the Matlab software on a personal computer (CPU i5-3230M 2.6 GHz and 4GB RAM).

Statistical type I error α suggests the probability of producing false alarms, i.e. detecting normal regions as flaws. Statistical type II error β implies the probability of producing missing alarms, which fail to alarm real flaws. We divide the area of normal region detected as flaws by the area of actual normal region to obtain type I error, and the area of undetected flaws by the area of actual flaws to obtain type II error. Therefore, the correct classification rate (CR) is defined as: $CR = (N_{cc} + N_{dd}) / N_{total}$ where N_{cc} is the pixel

number of normal textures detected as normal areas, N_{dd} is the pixel number of flaws detected as defective regions, and N_{total} is the total pixel number of a testing image.

To evaluate the impact of varying number of decomposition levels on the reconstruction result, Figures 5(a)-5(d) present the restoration results from decomposition levels 1, 2, 3, and 4, respectively. All these images are solely reconstructed from a filtered approximated sub-image and three corresponding detailed sub-images with the Haar wavelet. Both Figures 5(a) and 5(b) reveal that too small the number of decomposition levels (such as 1 and 2) cannot sufficiently separate defects from the repetitive texture patterns (false alarms). However, too large the number of decomposition levels (such as 4) yields the fusion effect of the defects, any may result in erroneous detection (missing alarms). The number of decomposition level 3 is most appropriate to enhance defects in the restored image. Experiments on a variety of textures images have confirmed that decomposition level 3 is generally sufficient for this area defect detection applications.

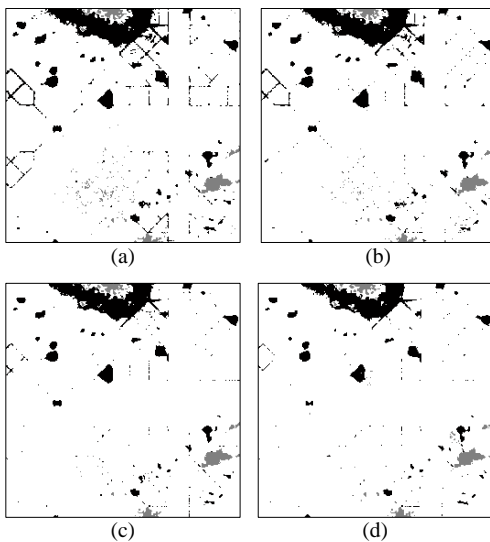


Figure 5. Resulting images of defect detection at four WT decomposition levels: (a) first decomposition level; (b) second decomposition level; (c) third decomposition level; (d) fourth decomposition level.

When various decision thresholds are used, their pairs of false alarm rates and detection rates are plotted as points on a Receiver Operating Characteristic (ROC) curve. The upper-left corner indicates a 100% detection rate and a 0% false alarm rate. The more the ROC curve approaches the upper-left corner, the better the test performs. In industrial practice, a more than 90% detection rate and a less than 10% false alarm rate are a good rule of thumb for performance evaluation of a vision system.

To completely filter out all homogeneous line patterns in the spatial domain image, both of the frequency

components on the principal bands and those frequency components in the neighborhood of the principal bands must be removed from the wavelet domain image. The filtering width determines the regions of the band neighborhoods will be filtered for high-energy frequency components. Figure 6 demonstrates the ROC curve of the proposed method with different flat zone sizes of k values, 0.7, 0.9, 1.1, 1.3, 1.5, respectively. It shows the defect detection performance of the proposed method with k value of 1.1 is better than those of the other k values. The flat zone filtering method with larger k value not only removes homogeneous line patterns but also local defects in the restored image and result in neglect small defects.

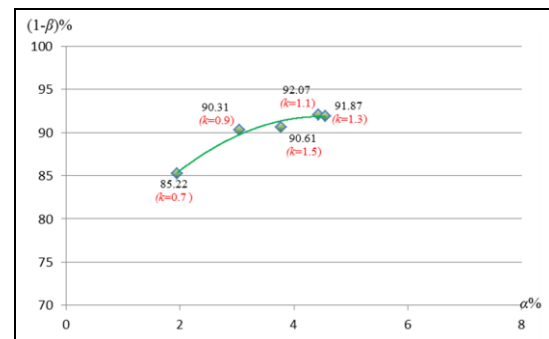


Figure 6. ROC plot of different flat zone sizes for filtering.

To demonstrate the flaw detection results, Figure 7 lists partial results of detecting area defects by the Otsu method [8], the Iterative method, the three level method, the proposed method, and the professional inspector, individually. The three spatial domain techniques, the Iterative, Otsu, and three level methods, make lots of erroneous judgments (false alarms) on area defect detection. The frequency domain technique, the proposed method, detects most of the area defects and makes less erroneous judgments. Therefore, the frequency domain approach outperforms the spatial domain techniques in the area defect detection of the touch panels.

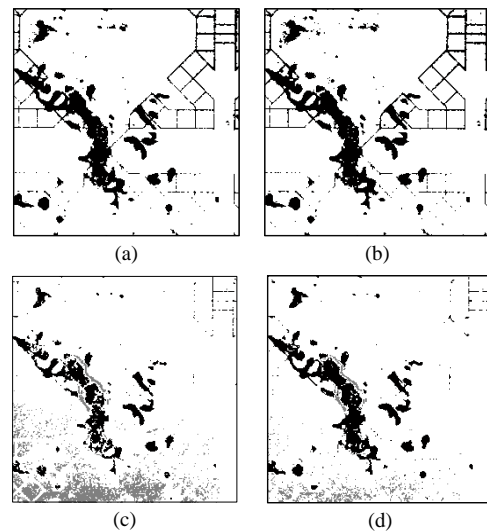


Figure 7. Partial detection results of Iterative, Otsu, three-level, and the proposed methods.

To compare the performance of the area defect detection, three spatial domain approaches and one frequency domain techniques are evaluated against the results by professional inspectors. The average defect detection rates ($1-\beta$) of all testing samples by the four methods are, respectively, 80.87% (Iterative method), 81.07% (Otsu method), 90.14% (three level method), and 91.82% (the proposed method). However, the three spatial domain methods have significantly higher false alarm rates (α), 20.41% (Iterative method), 26.70% (Otsu method), and 11.64% (three level method). On the contrary, the other frequency domain approach has rather lower false alarm rate, 4.36% (proposed method). Hence, the proposed method can overcome the difficulties of detecting area defects on touch panels and excels in its ability of correctly discriminating area defects from normal regions.

5 Conclusions

In this study, we have presented a frequency filtering approach for automated inspection of area defects in directional textures of touch panels. The line patterns of four directional textures in the spatial domain image can be easily removed by detecting the band region of an approximated sub-image of a decomposed image in wavelet domain, setting them to mean value by the flat zone filter, and transforming back to a spatial domain image. In the filtered restored image of a textured surface, the periodic line region in the original image will have an approximately uniform gray level. Experimental results show that the proposed method achieves a high 91.82% probability of correctly discriminating area defects from normal regions and a low 4.36% probability of erroneously detecting normal regions as defects on structural textured surfaces of touch panels.

6 Acknowledgment

This work was partially supported by the National Science Council (NSC) of Taiwan, under Grant No. NSC 101-2221-E-324-007-MY2.

7 References

- [1] H. D. Lin and H. H. Tsai, "Automated quality inspection of surface defects on touch panels," *Journal of the Chinese Institute of Industrial Engineers*, **29(5)**, 291-302 (2012).
- [2] H. Liu, Y. Wang, and F. Duan, "Glass bottle inspector based on machine vision," *International Journal of Computer Systems Science and Engineering*, **3:3**, 162-167, (2008).
- [3] E. N. Malamas, E. G. M. Petrakis, M. Zervakis, L. Petit and J. D. Legat, "A survey on industrial vision systems, applications and tools," *Image and Vision Computing*, **21**, 171-188 (2003).
- [4] A. Latif-Amet, A. Ertüzün and A. Ercil, "An efficient method for texture defect detection: sub-band domain co-occurrence matrices," *Image and Vision Computing*, **18**, 543-553 (2000).
- [5] C. S. Cho, B. M. Chung and M. J. Park, "Development of real-time vision-based fabric inspection system," *IEEE Transactions on Industrial Electronics*, **52**, 1073-1079 (2005).
- [6] D. B. Perng, W. C. Wang and S. H. Chen, "A novel quasi-contact lens auto-inspection system," *Journal of the Chinese Institute of Industrial Engineers*, **27**, 260-269 (2010).
- [7] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, **13** (1), 146-156 (2004).
- [8] N. Otsu, "A threshold selection method from gray level histogram," *IEEE Transactions on Systems, Man and Cybernetics*, **9**, 62-66 (1979).
- [9] H. F. Ng, "Automatic thresholding for defect detection," *Pattern Recognition Letters*, **27**, 1644-1649 (2006).
- [10] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, New Jersey (2008).
- [11] C. H. Chan and G. K. H. Pang, "Fabric defect detection by Fourier analysis," *IEEE Transactions on Industry Applications*, **36**, 1267-1276 (2000).
- [12] D. M. Tsai and B. Hsiao, "Automatic surface inspection using wavelet reconstruction," *Pattern Recognition*, **34**, 1285-1305 (2001).
- [13] H. D. Lin, "Automated visual inspection of ripple defects using wavelet characteristic based multivariate statistical approach," *Image and Vision Computing*, **25**, 1785-1801 (2007).
- [14] A. Kumar and K. H. Pang, "Defect detection in textured materials using Gabor filters," *IEEE Transactions on Industry Applications*, **38**, 425-440 (2002).
- [15] H. D. Lin and J. D. Jiang, "Applying discrete cosine transform and grey relational analysis to surface defect detection of LEDs," *Journal of the Chinese Institute of Industrial Engineers*, **24**, 458-467 (2007).
- [16] H. D. Lin, "Tiny surface defect inspection of electronic passive components using discrete cosine transform decomposition and cumulative sum techniques," *Image and Vision Computing*, **26**, 603-621 (2008).

- [17] D. M. Tsai and C. Y. Hsieh, "Automated surface inspection for directional textures," *Image and Vision Computing*, **18**, 49-62 (1999).
- [18] D. B. Perng and S. H. Chen, "Automatic surface inspection for directional textures using nonnegative matrix factorization," *International Journal of Advanced Manufacturing Technology*, **48**, 671-689 (2010).
- [19] Y. C. Chen, J. H. Yu, M. C. Xie and F. J. Shiou, "Automated optical inspection system for analogical resistance type touch panel," *International Journal of the Physical Sciences*, **6** (22), 5141-5152 (2011).
- [20] S. Arivazhagan and L. Ganesan, "Texture segmentation using wavelet transform," *Pattern Recognition Letters*, **24**, 3197-3203 (2003).
- [21] M. K. Bashar, T. Matsumoto, and N. Ohnishi, "Wavelet transform-based locally orderless images for texture segmentation," *Pattern Recognition Letters*, **24**, 2633-2650 (2003).
- [22] S. C. Kim and T. J. Kang, "Texture classification and segmentation using wavelet packet frame and Gaussian mixture model," *Pattern Recognition*, **40**, 1207-1221 (2007).
- [23] W. C. Li and D. M. Tsai, "Wavelet-based defect detection in solar wafer images with inhomogeneous texture," *Pattern Recognition*, **45**, 742-756 (2012).
- [24] D. M. Tsai and C. H. Chiang, "Automated band selection for wavelet reconstruction in the application of defect detection," *Image and Vision Computing*, **21**, 413-431 (2003).

Traffic Control by Digital Imaging Cameras

Rowa'a Jamal, Karmel Manaa, Maram Rabee'a, Loay Khalaf
Electrical Engineering Department, University of Jordan, Amman, Jordan

Abstract— *Traffic Control is considered one of the fastest developing technologies in the world. One such method is control by traffic cameras. The first cameras installed for traffic monitoring were developed in the 1960s. This development led to the growth of multi-purpose traffic cameras in several countries across the world. The scope of this report is concentrated on producing a traffic control camera, which can be installed at crossroads with traffic lights. Algorithms for speed detection, and license plate recognition, are described and their performance is evaluated.*

Keywords— Image processing, traffic camera, plate's recognition, traffic.

1 Introduction

Road crashes are considered as one of the major causes of death and injury. Over the years, there has been a noticeable steady increase of traffic violations and problems. This led the search for ways to control traffic with the intentions of reducing collisions by enforcing traffic laws, including traffic light violation, red light violation, and speed violation. Over the last couple of decades, researchers have deliberately worked on improving the control at traffic intersections and traffic lights to reduce traffic jams and accidents. The bottleneck in traffic problems is related to the limited resources in the current infrastructure, such as two roads crossing or merging. The traffic problems get worse with time, since the number of vehicles is increasing significantly.

The use of automated traffic control technologies is now wide spread throughout the world. Worldwide, despite of variation in the nature of these applications, they have provided positive road safety benefits. The first red light camera was introduced in 1965 in the Netherlands. This camera was based on using tubes stretched across the road to detect the violation and subsequently trigger the camera. [1]. New York's red-light camera program went into effect in 1993. [2].

The first digital camera system was introduced in [Canberra](#) in December 2000, and digital cameras have increasingly replaced the old film cameras in other locations since then [3]. From the late 1990s, digital cameras began to be introduced, those cameras can be installed with a network connection to transfer real time live images to a central processing unit, for that

they have advantages over film cameras in speed monitoring. However, film-based systems may provide superior image quality in the variety of lighting conditions encountered on roads. New film-based systems are still being used, but digital ones are providing greater proficiency, lower maintenance and are now more popular.

2 Paper Overview

The paper discusses the production of a traffic control camera used to obtain red light violation, license plate recognition (LPR), and speed detection of the vehicles. The proposed camera is designed to be used at the bottleneck of traffic; intersections with traffic lights. Since there are several technologies used to obtain the same aim of this study, other characteristics were taken into consideration to make it more attractive.

The major step in using such cameras is monitoring the traffic at the red light by capturing a video. Then the video is processed by using image processing techniques. The image processing code was developed using Matlab 7.7, whereby the program reads video file, convert it to frames, and then by character segmentation, it can recognize the type of the violation. The main tasks of this camera include detection of the red light violation. Simply, the camera will check the color of the light, if it was red, it will compare between a saved picture for the street in such red light case (the street in front of the traffic light must be empty) and the captured one. If there is any violation, the camera will capture a photo for the car and perform plate recognition.

On the other hand, for the speed violation, the camera will measure the distance between two points passed by the car and the time elapsed through this distance. Then, it will calculate the speed of the car by dividing the distance over the time. Also, if there is any violation in the speed, it will capture and perform plate recognition for the car. The system continuously monitors the traffic signal and the camera is triggered by any vehicle entering the intersection. Automatic number plate recognition can be used for purposes unrelated to enforcement of traffic laws. In principle, any agency or person with access to data either from traffic cameras or cameras installed for other purposes can track the movement of vehicles for any purpose.

3 Implementation

The main effort of this research is the use of image processing of captured images of a digital camera. The images used in this project and the video were taken by an inexpensive Canon 500 D, digital camera of 15 megapixel and 3.00x zoom (as shown in the figure below). It records videos with very high accuracy, full HD with 30 frames per second. Throughout this section, the design essentials, basics and procedure will be discussed separately.



Figure 1: Canon 500 D Digital Camera

The desired system should be able to meet the requirements and goals stated below:

- Ability to detect the speed of the vehicle that crosses the traffic light with accuracy of 15%.
- Ability to recognize the vehicle registration plate with accuracy of 75%.
- Ability to identify cars crossing the red light.

4 Traffic Detectors

This section describes current technology associated with traffic control.

4.1 Induction Loops

Inductive Loop Detector technology has been in use for the detection of vehicles since the early 1960's. It consists of a loop of wire and an electronic detection unit. Simply, the operation is based on the principle of metal detection, relying on the fact that a moving metal will induce an electrical current in a nearby conducting wire. With a vehicle detector, the loop is buried in the roadway and the object to be detected is a vehicle. (As shown in the figure below).

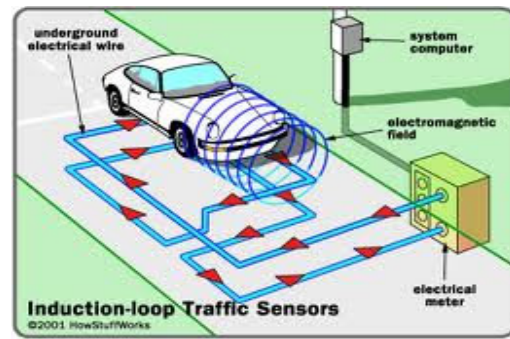


Figure 2: Induction Loop

Vehicle detection loops are used to detect vehicles passing a certain area, for our approach; a traffic light. An insulated, electrically conducting loop is installed in the pavement. The electronic unit transmits energy into the wire loops at frequencies between 10 kHz to 200 kHz, depending on the model. The inductive-loop system behaves as a tuned electrical circuit in which the loop wire is considered as the inductive elements. When a vehicle passes over the loop or is stopped over the loop, the vehicle induces eddy currents in the wire loops, which decrease their inductance. The decreased inductance actuates the electronics unit output relay or solid-state optically isolated output, which sends a pulse to the traffic signal controller signifying the passage or presence of a vehicle.

4.2 Microwave Radar

Radar is an object-detection system which uses [radio waves](#) to determine the range, direction, or speed of objects. It can be used to detect [motor vehicles](#). The radar antenna transmits pulses of radio waves which bounce off any object in their path. The object returns a small part of the wave's energy to the receiver antenna which is usually located at the same site as the transmitter.

The basic use of the traffic radars is the measurement of the speed of the vehicle. Traffic radar calculates speed from the reflections it receives. It uses a phenomenon of physics known as the Doppler principle. The classic example is heard along roads. As the car approaches, you hear the high pitch sound of the car horn. The instant the car passes and begins to move away, you hear a lower pitch sound. The car itself is making the same sound both coming and going, but for a stationary listener, the speed of the car adds to the pitch of its sound as it approaches, and subtracts as it departs. This change from true pitch is called the Doppler shift, and the magnitude of the change depends upon the speed of the car. The Radar compares the shifted frequency of the reflection to the original frequency of the beam it sent out and from the difference it calculates speed. The figure below shows how microwave radar detects the speed of the vehicle.

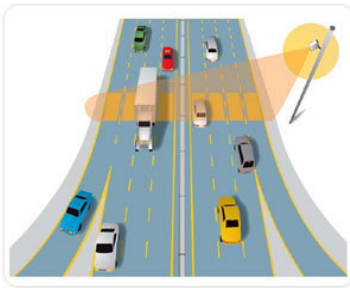


Figure 3: Microwave Radar

4.3 Infrared Sensors

Active and passive infrared sensors are manufactured for traffic flow monitoring applications. Active infrared sensors illuminate detection zones with low power infrared energy transmitted by laser diodes operating in the near infrared region of the electromagnetic spectrum. A portion of the transmitted energy is reflected or scattered by vehicles back towards the sensor. The diodes operated in the near infrared spectrum at 880 nanometers (nm). The signal modulation prevented interference from other sources of infrared energy, including sunlight. Two transmitter-receiver systems measured the vehicle speed and one measured the vehicle height. When trucks susceptible to rollover or jackknifing were encountered, flashers were activated to warn drivers to reduce speed. [4].

Passive sensors do not transmit energy; they detect energy from two sources:

- 1- Energy emitted from vehicles, road surfaces, and other objects in their field-of-view.
- 2- Energy emitted by the atmosphere and reflected by vehicles, road surfaces, or other objects into the sensor aperture

The energy captured by infrared sensors is focused by an optical system onto an infrared-sensitive material mounted at the focal plane of the optics. This material converts the reflected and emitted energy into electrical signals. Real-time signal processing is used to analyze the signals for the presence of a vehicle. The sensors are mounted overhead to view approaching or departing traffic. They can also be mounted in a side-looking configuration. Infrared sensors are utilized for signal control; volume, speed, and class measurement; detection of pedestrians in crosswalks; and transmission of traffic information to motorists.

4.4 Video Detection

Video detection is based on real-time image processing providing efficient wide-area detection well suited for registration of incidents on roads and in tunnels. Connected to Traffic Controllers, the application can also be used for vehicle detection at signalized intersections where it is difficult or expensive to install inductive loops. Video-detection systems are also considered non-intrusive.

Video detection combines real-time image processing and computerized pattern recognition in a flexible platform; it uses a vision processor to analyze real-time changes in the image. In

this system, cameras called image sensors capture images and provide a video signal to the vision processor. The video signal is analyzed and the results are recorded. Video image detection is one of the leading alternatives to the commonly used loop detectors. It is progressively being used to detect traffic intersections and interchanges. This is because video detection is often cheaper to install and maintain than inductive loop detectors at multi-lane intersections. In addition to speed, volume, queues and headways, it provides traffic engineers with many other traffic characteristics, such as level of service (LOS), space mean speed, acceleration and density. Video detection is also more readily adaptable to changing conditions at intersections (e.g., lane reassignment and temporary lane closure for work zone activities). This is one of the biggest advantages of video image detection. It provides traffic managers with the means to reduce congestion and improve roadway planning. Additionally, it is used to automatically detect incidents in tunnels and on freeways, thus providing information to improve emergency response times of local authorities [5].

Through the discussion about the image-processing cameras, it is noticeable that they have these advantages:

- Monitors multiple lanes and multiple detection zones/lane.
- Easy to add and modify detection zones.
- Rich array of data available.
- Provides wide-area detection when information gathered at one camera location can be linked to another.
- Generally cost-effective when many detection zones within the camera field-of-view or specialized data are required.

5 Image Processing

Image processing is defined as a process involving the change of the natural appearance of an image. It consists of an input and an output. The input is an image whereas the output is a set of characteristics related to the image, also the output maybe an image.

The main aim of image processing lies in converting the image for better human interpretation and machine perception. The operation of image processing may contain several actions including making the images to appear sharper, removing motion blur from images, removing noise from images, obtaining the edges of an image and removing details from an image [6].

5.1 Basic Types of Images

There are a total of four basic types of images, namely Binary, Grayscale, True color or red-green-blue, and Indexed. The descriptions of all these images are as follows:

5.1.1 Binary Image

A binary image is a digital image that has only two possible values for each pixel. The pixel is made up of either in black or

white color. Binary images are also called bi-level or two levels. This means that each pixel can be stored as a single bit in either binary '0' or '1'. Figure 4 shows a binary image.

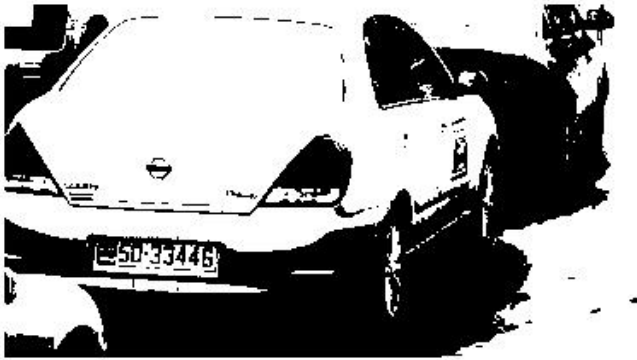


Figure 4: Binary Image

5.1.2 Grayscale Image

A grayscale digital image is an image in which the value of each pixel is not the same. The only colors are shades of gray. Each image pixel is made up of shade of gray from 0 (black) to 255 (white). Each pixel can be represented by 1 byte or 8 bits. The reason for having such an image was because less information is needed to be provided for each pixel.



Figure 5: Grayscale Image

5.1.3 True colour or red-green-blue (RGB) Image

Each pixel will have a particular color that is being described by the amount of red, green, and blue in it. If each of the components has a range of 0-255, this will give a total of 2^{24} different possible colors in an image. Each pixel will require 24 bits and they are called 24-bit color images.



Figure 6: RGB Image

5.1.4 Indexed Images

Each pixel has a value that does not give its color but an index to the color in the map which has a list of all the colors used in that image.

6 Project Design

The procedure followed to obtain the interruption whether for the red violation or the speed violation is shown in Figure 6.

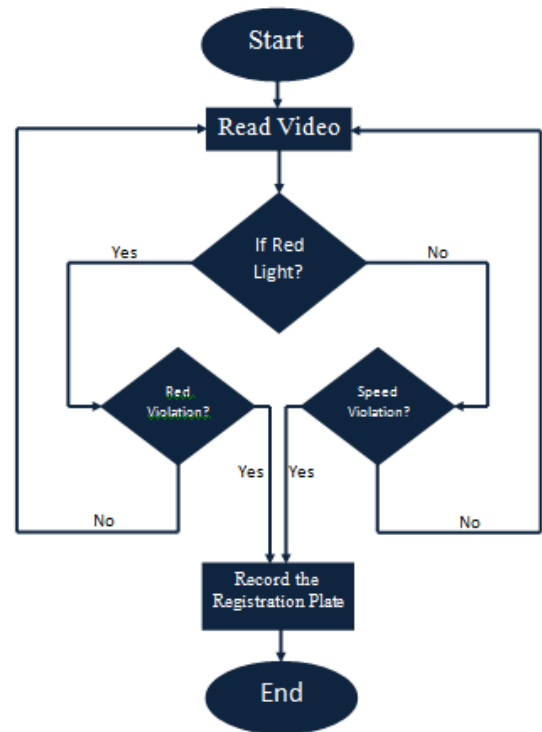


Figure 7: System Flow Chart

6.1 Red Light Violation

This part has been made by taking the absolute difference between the current inputs with a reference one. So it could be known if the car had passed the red line.

By applying the following steps:

- Define the red light position in the traffic light.
- Define the reference frame to compare with (when the street is empty).
- Define the frame where the violation occurred.
- Take the absolute difference between the two frames.

For that it could be known if the car had passed the red line showed in the figure below:

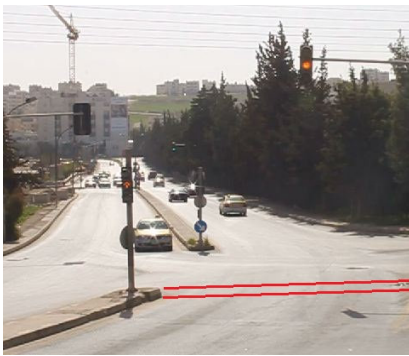


Figure 8: Reference of Red Light Violation

6.2 Speed Violation

This part is to determine if the car has crossed a certain speed limit which was chosen to be 600 pixel/second for normal traffic conditions. Two reference points have been defined $c1=[162\ 260]$, $c2=[445\ 504]$, $r1=[691\ 700]$ and $r2=[567\ 577]$ (as shown in the figure below). This operation reads the video frames and first if there was a crossing found in them, it takes the difference which is the pixels difference between the points. The absolute difference is taken between the reference defined by (c1 and r1) and the read frame. The difference is defined as start frame, then it takes another absolute difference between (c2,r2) and the read frame, where the difference is defined as the final frame.

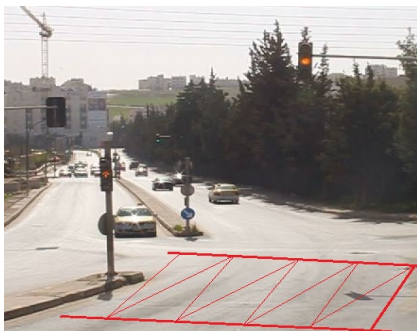


Figure 9: Speed Violation

The difference is made by:

$$\mathit{Framdiff} = \mathit{finalfram} - \mathit{startfram}; \quad (1)$$

The time can be known from the bellow equation were 30 is the frame rate:

$$\mathit{Time} = \frac{\mathit{framdiff}}{30}; \quad (2)$$

Then the speed is calculated by the equation bellow, where 310 is the difference in pixels:

$$\mathit{Speed} = \frac{310}{\mathit{time}}; \quad (3)$$

6.3 Plate Numbers Recognition

This is done by first determining the plate position, then taking the numbers from the plate and from a look-up table to compare. The numbers are detected by taking the ones that have the maximum correlations. The look up table is a template that contains the numbers from 0 to 9.

The action of recognition includes many functions to obtain the exact result. The image will be transformed to Grayscale. Then the extended-maxima transform is computed. Edge detection and Strel functions are also used. Finally, plate shaping and filter processing are used to obtain the final result.

Plate Position Determination is made by the following steps:

- Transform the image to Grayscale.
- Computes the extended-maxima transform.
- Apply edge detection type Sobel.
- Determine the line by Strel function.
- Dilate the image.
- Verify the rectangular plate shape.

Plate Numbers Recognition consists of three parts:

- PNR(plate number recognition): in this M-file the image is being read and then opens a text document where the result is being displayed at. In this M-file the plate numbers are being divided each separately in order to the compare. As shown in figure 3.11.
- Creatt: in this M-file the templates has been made.
- Recognize_num: in this M-file it computes the correlation between template and input image and takes the maximum correlation. Its output is a string containing the letter.

The correlation is made by this part of the code:

$$\mathit{for\ } n = 1:\mathit{all_numb} \quad (4)$$

$$\% \mathit{where\ } \mathit{all_numb} \mathit{ is\ from\ the\ PNR\ M - file\ which\ is\ the\ size\ templates} \\ \mathit{corr} = \mathit{corr2}(\mathit{templates}\{1, n\}, \mathit{cropped}); \quad (5)$$

$$\% \mathit{this\ computes\ the\ correlation\ between\ the\ templates\ and\ the\ input\ image.} \\ \mathit{corr_values} = [\mathit{corr_value}\ \mathit{corr}]; \quad (6)$$

After applying this part of the project the result of the plate number will appear in a text document.

The templates that were used in finding the maximum correlation:



Figure 10: Templates for the Plate Numbers

55555	55555	0%	100%
11-8880	498880	42.86%	57.14%
11-430	1-433	33.33%	66.67%
11-1111	4941411	57.14%	42.86%
10-962	10-862	16.67%	83.33%
Total Accuracy=			70.38%

7 Performance Analysis

The proposed design was tested at 25 plates; the accuracy percentage was determined by the following:

$$\text{Accuracy percentage} = \frac{\text{number of matching numbers}}{\text{total number}} \quad (7)$$

Note: any addition number appear in the result count as an error, the table below shows the results and the percentage of error:

Table 1: Results Summary

Plate number	Result	Error percentage	Accuracy percentage
22-4444	22-4444	0%	100%
2-70000	2 70000	0%	100%
4-44444	4-44444	0%	100%
14-17772	-147772	25%	75%
34-444	34444	0%	83.33%
15-55000	1555000	12.5%	87.5%
10-5000	10-5000	0%	100%
10	7952	40%	60%
952	2162	25%	75%
60-13583	80013583	25%	75%
71-6516	75685-8	71.43%	28.57%
18	809614	37.5%	62.5%
80964	157754742895	50%	50%
15-74995	157754742895	50%	50%
22-95951	22185851	37.5%	62.5%
7-6628	748828	50%	50%
7-3052	743052	16.67%	83.33%
71-9810	71488-10	37.5%	62.5%
187687	167667	33.33%	66.67%
20-67120	2087120	25%	75%
20-74717	23	87.5%	12.5%

The figure below represents the PNR result:

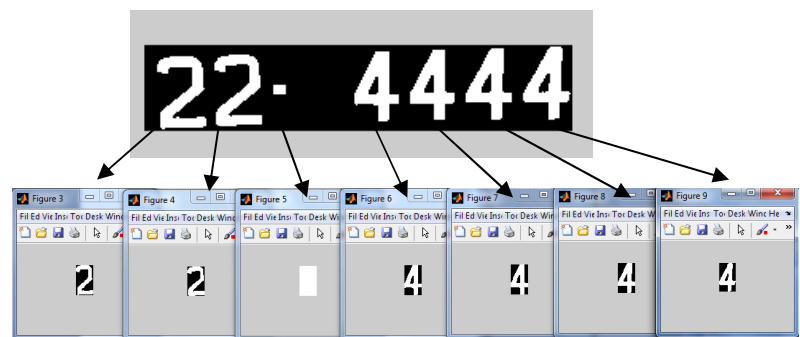


Figure 11: PNR

The figure below represent the correlation between the input numbers and the templates, in the illustrated plate, the figure shows the correlation for the first element number two and the templates the figure shows that it's maximum correlation was with templates of number two:

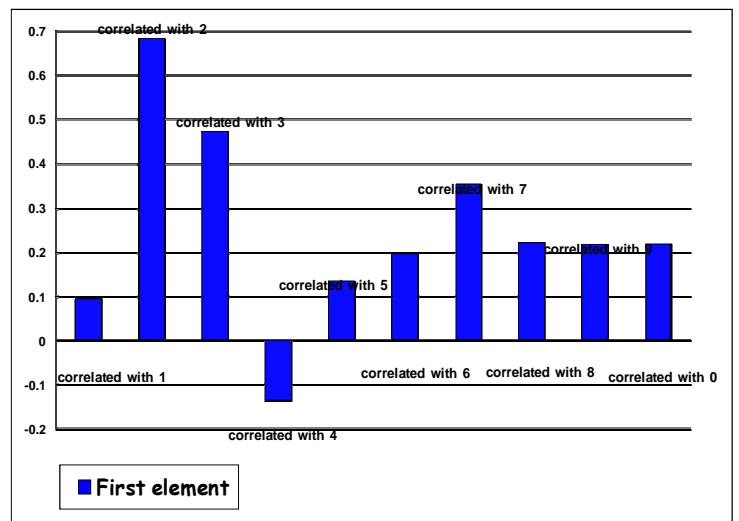


Figure 12: Correlation of Number (2)

8 General Conclusion

The results of image processing were very good for the purpose of this project; the operations used to edit the images have done the task successfully. The camera specifications played an important role in the accuracy of the code, as it is better, the accuracy will be better.

This traffic control system can make a good drop in the cost comparing to the other systems are being used in the traffic controlling, as it only needs a computer, and a high quality digital camera.

In the plate recognition what we faced was that some of the plates have some nails on them or some stains which make it harder to extract the numbers correctly.

The same approach that was used in this project can be used to detect the traffic congestion in a street and therefore control the traffic light signaling depending on the numbers of cars in the street, which can reduce the congestion on the streets.

9 References

- [1] Retting, Richard A.; Ferguson, Susan A.; Hakkert, A. Shalom, 2003, "Effects of Red Light Cameras on Violations and Crashes: A Review of the International Literature".
- [2] D.J. Dailey and L. Li, Apr 2000, "VIDEO IMAGE PROCESSING TO CREATE A SPEED SENSOR".
- [3] Peter Clack, Nov 26, 2000, "World-First Digital Camera to Nab Red Light Runners".
- [4] Ralph Gillman "Office of Highway Policy Information"
- [5] <http://www.mountain-plains.org/pubs/html/mpc-04-166/pg1.php> - May - 2012
- [6] "Matlab tutorial fundamental programming."

An Orange Sorting Technique based on Size and External Defects

Naeem Sattar, Sheikh Ziauddin, Sajida Kalsoom, Ahmad R. Shahid, Rafi Ullah, Amir H. Dar

Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Abstract—In this paper, a new automated orange sorting technique is presented. It sorts orange fruit based on size and external defects. We use image processing techniques for segmentation of oranges and then based on our algorithm categorize them according to the presence and type of defect. The defects considered are anthracnose, unripe and stem-end injury. Non-defected oranges are further categorized into large, medium or small based upon their size. We created an image dataset containing images of 189 oranges. We achieve high accuracy with success rates of 97.4% and 100% for defect-based and size-based sorting, respectively.

Index Terms—Precision agriculture, citrus grading, citrus sorting, computer vision in agriculture, image processing in agriculture

I. INTRODUCTION

Orange is a significant horticultural crop with annual world wide production of more than 50 million tonnes each year [1]. According to United States Department of Agriculture, Brazil has the largest orange fruit production in the world (17.8 million metric ton) followed by China, United States and the European Union [1].

The homogeneity and presentation of fruit has considerable effect on the customers' decision to buy fruit. It can be improved with better packaging of fruit, and using better transportation and marketing. The task of packaging, which includes sorting and grading of fruit, can be carried out using manual labor which is not only prone to errors but it also incurs extra cost in terms of labor hired. Therefore, use of fast and accurate vision methods for sorting and grading is highly desirable.

Citrus fruit can be categorized by its physical characteristics, such as shape, size, maturity, volume, color and defects [2]. In this paper, we have categorised orange fruit based on size and external defects, i.e. *large*, *medium* and *small* for size; *anthracnose*, *unripe* and *stem-end injured* for defects. We have used image processing techniques to accurately classify the fruit. As

a first step, red channel analysis is performed to separate non-defected or stem-end injured from anthracnose or unripe fruit. The former are separated by calculating the area while the latter are classified using hue and saturation values. Finally, non-defected oranges are categorised based on their sizes into *small*, *medium* and *large* categories.

The rest of this paper is organised as follows. Section II presents an overview of some of the important existing schemes for fruit grading. Section III describes our algorithms for defect-based and size-based orange sorting. Experimental results are presented in Section IV while we present conclusion of this work in Section V.

II. RELATED WORK

Image processing and computer vision based fruit classification schemes have been presented for diverse horticultural crops such as orange, mango, apple, lemon and pomegranate [3] [4] [5] [6] [7]. Leeman et al. [8] presented an algorithm to categorize apple fruit based on their external defects into extra, I, II and reject category. They selected two classes of apples for their experiments, i.e. Golden Delicious and Jonagold apples. Using Fishers linear discriminant analysis, neural networks and a correlation pattern recognition technique [9], they managed to achieve success rates of 78% and 72%, for Golden Delicious and Jonagold apples, respectively.

Koc [10] used ellipsoid approximation and image processing disk approximation methods to estimate the volume of watermelons. They compared them with the volume calculated using traditional water displacement method. They showed that the volume calculated using image processing technique provides a good agreement with the volume calculated using water displacement method as compared to the volume computed using ellipsoid approximation.

Textural and Geometrical features have also been used to categorise citrus fruit. Aleixos et al. [11] presented

an algorithm to categorize citrus fruit based on size, colour and external defects along with the development of parallel hardware and software architecture. A quality inspecting technique for oranges, peaches and apples was presented by Blasco et al. [12] where fruits were graded based on their size, colour, stem location and external defects. Bayesian discriminant analysis was used for segmentation process then morphological features were extracted, e.g., fruit size and centroid, followed by the external features, such as length of the major damage, the damaged area, the stem and calyx, and the primary and secondary colour. After testing the system in real environment, its precision and repeatability were found to be similar to those of manual grading.

Blasco et al. [13] categorized oranges and mandarins based on their defects. They analysed five different colour spaces. In addition, near-infrared, ultraviolet and fluorescence imaging were used to improve the quality of classification. They combined the spectral information and showed capability of identifying most defects with high accuracy. In subsequent work, Blasco et al. [14], [15], used morphological operators to further enhance the performance of multispectral imaging. They tested their system for automatic sorting of satsuma segments in the real time environment.

Another technique to measure the volume and mass of four different citrus fruits; lemons, limes, oranges, and tangerines was presented by Omid et al. [16]. They used elementary elliptical frustums to estimate the fruit volume which in turn was used to estimate the mass of the fruit.

III. THE PROPOSED SCHEME

Our proposed scheme has two phases: image segmentation and orange sorting. In segmentation phase, we perform some preprocessing steps to extract the region of interest. The segmented orange image is then processed to identify different defects. We have classified oranges as either *non-defected* or having one of the following three defects: *anthracnose*, *unripe* and *stem-end injury*. Figure 1 shows one image from each of these classes.

The fruit which is identified as non-defected is next processed to find its size. We have classified fruit into three sizes: *large*, *medium* and *small*. The decision about the size category of the fruit is done manually and then compared against the one classified by our algorithm. Figure 2 shows a graphical depiction of different steps and decisions involved in our algorithms.

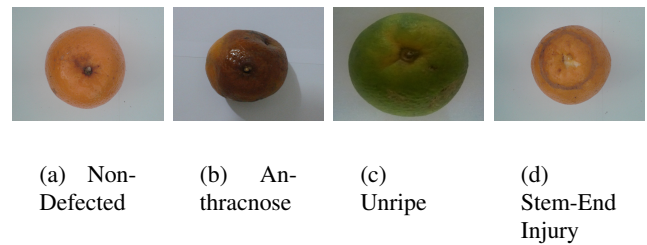


Fig. 1. Sample images from our dataset, a) A non-defected orange, b) An orange having anthracnose, c) An unripe orange, d) An orange having stem-end injury.

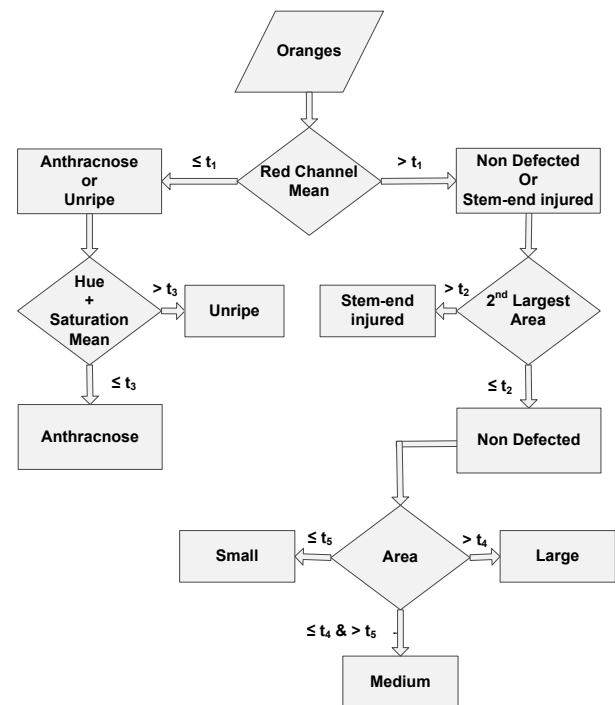


Fig. 2. A flowchart showing the working involved in the proposed scheme.

A. Orange Segmentation

Segmentation of oranges within the images is performed in this step. Firstly, we convert the input image into HSV form and generate binary masks for hue and saturation channels. These masks are generated by applying adaptive thresholding technique of Otsu [17] on hue and saturation bands separately. In Otsu's technique, a global threshold is computed which converts an intensity image into a binary image. The idea is to maximize the inter-class variance by minimizing the intra-class variance between the foreground and the background pixels. The weighted intra-class variance $\sigma_w^2(t)$ is given

by

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \quad (1)$$

where t represents the desired threshold, $q_1(t)$ and $q_2(t)$ are probabilities, and $\sigma_1^2(t)$ and $\sigma_2^2(t)$ are variances of background and foreground pixels, respectively. The class probabilities are computed as

$$q_1(t) = \sum_{i=1}^t P(i) \quad (2)$$

$$q_2(t) = \sum_{i=t+1}^L P(i) \quad (3)$$

where $P(i)$ is the probability of intensity level i and L is the number of total intensity levels. The class variances are given by

$$\sigma_1^2(t) = \sum_{i=1}^t (i - \mu_1(t))^2 \frac{P(i)}{q_1(t)} \quad (4)$$

$$\sigma_2^2(t) = \sum_{i=t+1}^L (i - \mu_2(t))^2 \frac{P(i)}{q_2(t)} \quad (5)$$

where $\mu_1(t)$ and $\mu_2(t)$ are means of background and foreground pixels, respectively.

After we obtain the hue and the saturation masks, we combine (using logical AND operation) these masks to get a single mask. This resultant mask is a binary image representing the fruit as white pixels and background as black pixels. Although this mask is a reasonable representation of the segmented orange but still there is some noise present which we remove by applying different morphological operations. First we apply morphological opening on the mask which has the effect of removing small connecting components which is followed by a morphological closing operation which has the effect of removing any small holes present in the mask. Finally, we apply this mask on the original RGB image. A combination of all these operations have the effect of removing the background (which is mostly of light color) and segmenting the foreground orange fruit. Figure 3 shows the output of intermediate steps involved in the segmentation algorithm.

B. Defect-Based Orange Sorting

The segmented image is then processed for defect-based sorting. We have four classes here: non-defected, anthracnose, unripe and stem-end injury. For each image, we find the mean of its red channel values in the region of interest (ROI). Here ROI is the part which contains

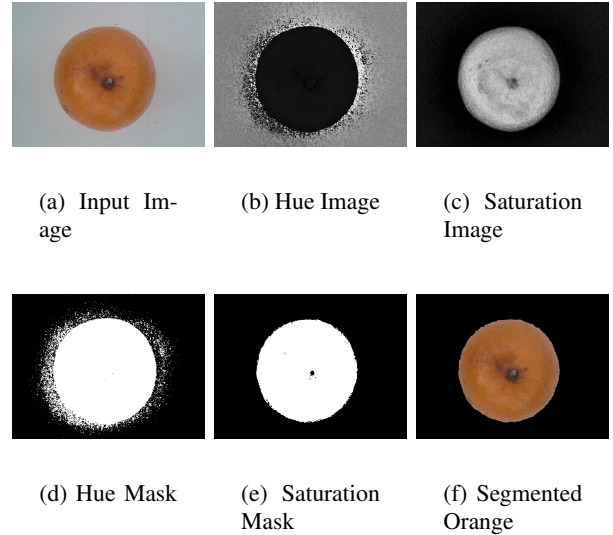


Fig. 3. Illustration of our orange segmentation algorithm's processing.

the segmented orange as shown in Figure 3(f). The red mean value (RMV) is given by

$$RMV = \frac{1}{n} \sum_{i=1}^n R(i) \quad (6)$$

where $R(i)$ is the red channel value at i th index of ROI and n represents the total pixels in ROI. Next, we compare the calculated RMV against a predefined threshold value. If the input value is greater than the threshold, we consider it to be either non-defected or having stem-end injury. On the other hand, if the input value is less than the threshold, we classify it to be either unripe or having anthracnose (see Figure 2). The intuition here is that the normal fruit as well as that with stem-end injury (with no other defect) has a bright orange color resulting in higher red channel mean values. On the other hand, the fruit having anthracnose is generally darker and unripe fruit is generally greenish so they tend to have lower red channel values.

Till this stage, we have classified orange fruit into two super-classes: 1) Non-defected or stem-end injured, and 2) Anthracnose or unripe. Next, we differentiate within these super-classes to achieve our desired classification.

To differentiate between non-defected and stem-end injured fruit, we adopt the following technique. First, we convert RGB image into grayscale and then to binary using Otsu's adaptive thresholding method [17]. Next, from this binary image, we find the areas of the largest object and the second largest object, respectively. Say these areas are A_1 and A_2 , respectively. Then we find

the ratio A_2/A_1 and apply a threshold value to this calculated ratio. If the ratio is greater than the threshold, we classify it as stem-end injury otherwise we classify it as non-defected. The reason of using this technique is based on the observation that, in stem-end injured fruit, a reasonably large portion near the stem is removed along with the stem. This portion is detected as a different object from the larger fruit object resulting in an A_2/A_1 value which is reasonably large. On the other hand, non-defected fruit either has no such second object (A_2/A_1 is 0) or has such an object but it is very small (A_2/A_1 is quite small).

To differentiate between anthracnose and unripe fruit, we use HSV image. We add hue and saturation band values in the region of interest and find their mean for each image. We notice that this mean value is generally higher for unripe fruit as compared to the fruit having anthracnose. As a result, we apply a simple threshold on this mean value to differentiate between anthracnose and unripe oranges.

C. Size-Based Orange Sorting

In the last section, we have applied multiple techniques to classify oranges correctly based on different defects (or no defect). If a fruit is classified as non-defected by the above technique, then we pass it to a size filter to classify it according to its size which we have categorized into three classes namely large, medium and small.

For size-categorization, we adopt a simple yet effective technique. First, we convert the image into grayscale and find the corresponding edge map. For finding an optimal edge map, we tested multiple edge detectors with different parameters and found that the Canny detector [18] provided best empirical results in terms of size-based classification. Canny edge detector first blurs the image using a Gaussian filter to reduce noise. Then the gradient is calculated to find variations in intensity values. After the gradient calculation, edge thinning is performed using non-maxima suppression. At the end, hysteresis thresholding is applied using a low and a high threshold to suppress those weak edges which are not connected to the strong ones.

After the edge image is calculated, we find the minimum bounding rectangle for the image which is the smallest rectangle that contains all foreground pixels. The length and width of minimum bounding rectangle gives us the horizontal and the vertical diameters of the object (orange). We take the mean of horizontal and vertical diameter which gives us the orange diameter.

TABLE I
NUMBER OF DEFECTED AND NON-DEFECTED FRUIT IMAGES IN OUR DATASET.

Non-Defected				Defected			
Small	Medium	Large	Total	Unripe	Anthracnose	Stem-End Injury	Total
30	30	30	90	25	52	22	99

TABLE II
CONFUSION MATRIX REPRESENTING OUR RESULTS FOR DETECTION OF DIFFERENT DEFECTS IN ORANGE FRUIT.

Detected Defect	Actual Defect			
	Unripe	Anthracnose	Stem-End Injury	Non-Defected
Unripe	25	0	0	0
Anthracnose	0	50	0	0
Stem-End Injury	0	0	20	1
Non-Defected	0	2	2	89
Total	25	52	22	90
Success Rate (%)	100	96.2	90.9	98.9

Using this diameter, we calculate the area of the object. We set two thresholds; if the calculated area is greater than the larger threshold, we classify the orange as large; if the area is lesser than the smaller threshold, we categorize it as small and finally, if the calculated size lies between the smaller and the larger thresholds, we consider the fruit as medium sized.

IV. EXPERIMENTAL EVALUATION

A. Orange Dataset

We created our orange fruit dataset by purchasing fruit from the local market. Our dataset contains a total of 189 images. 90 oranges are non-defected whereas the remaining 99 contain one of the three defects. Out of defected fruit, there are 52, 25 and 22 fruits for anthracnose, unripe and stem-end injury, respectively. For non-defected oranges, there are 30 oranges each for large, medium and small categories. Table I gives a tabular list of our complete dataset.

B. Defect-Based Sorting Results

We applied our algorithm to all dataset images and evaluated the accuracy for classifying oranges into multiple defect classes. Table II summarizes our results.

TABLE III
CONFUSION MATRIX REPRESENTING OUR RESULTS FOR
DETECTION OF DIFFERENT SIZES OF ORANGE FRUIT.

Detected Size	Actual Size		
	Small	Medium	Large
Small	30	0	0
Medium	0	30	0
Large	0	0	30
Total	30	30	30
Success Rate (%)	100	100	100

As it can be seen from the table, 100% accuracy was achieved in classification of unripe fruit followed by accuracies of 96.2%, 90.9% and 98.9% for anthracnose, stem-end injury and non-defected fruit, respectively. Overall, the algorithm was able to correctly classify 184 out of 189 fruits yielding an average success rate of 97.4%. We note that although anthracnose and stem-end injury show comparatively lower accuracy but still their accuracy is quite high. For misclassification in stem-end injury category, one reason is that both misclassified images were quite challenging in the sense that they did not show any significant signs of skin removal near the stem. As a result, the algorithm could not find a large enough value of A_2 (Section III-B) resulting in a wrong classification.

C. Size-Based Sorting Results

We applied our algorithm to all non-defected images and evaluated the accuracy for classifying oranges into large, medium or small classes. Table III summarizes our results. As we can see, the system is able to classify fruit with 100% accuracy.

D. Comparison with Existing Techniques

In this section, we compare our results with some of the existing techniques. Most of the existing work either perform defect-based grading or size-based grading. On the other hand, we have performed both types of gradings. Table IV compares our scheme with existing schemes in terms of success rates. As we see from Table IV, the results obtained by our algorithms are at par with the existing schemes.

V. CONCLUSION

In this paper, we propose and evaluate a scheme for accurately classifying orange fruit images acquired for orange sorting and grading systems. As opposed to most

TABLE IV
COMPARISON OF THE PROPOSED SCHEME WITH THE EXISTING
WORK.

Technique	Success Rate	
	Defect-Based Sorting	Size-Based Sorting
Blasco et al. [13]	95%	N/A
Blasco et al. [14]	86%	N/A
Aleixos et al. [11]	94%	N/A
Omid et al. [16]	N/A	94%
Proposed Scheme	97.4%	100%

existing schemes, which either use fruit size or external defects for this purpose, we use a technique considering both these factors for orange sorting. Experiments on 90 non-defected and 99 defected orange fruits show that the proposed technique achieves very high classification accuracy. In the future, we plan to extend the scheme by including more defect classes in the dataset. In addition, we plan to measure and compare fruit size in numeric values instead of large, medium and small classes in order to minimize the involvement of human error.

REFERENCES

- [1] United States Department of Agriculture, "Citrus: World markets and trade," apps.fas.usda.gov/psdonline/circulars/citrus.pdf, 2014, Online; accessed 15-March-2014.
- [2] Yud-Ren Chen, Kuanglin Chao, and Moon S Kim, "Machine vision technology for agricultural applications," *Computers and electronics in Agriculture*, vol. 36, no. 2, pp. 173–191, 2002.
- [3] KF Sanders, "Orange harvesting systems review," *Biosystems Engineering*, vol. 90, no. 2, pp. 115–125, 2005.
- [4] AB Payne, KB Walsh, PP Subedi, and D Jarvis, "Estimation of mango crop yield using image analysis–segmentation method," *Computers and Electronics in Agriculture*, vol. 91, pp. 57–64, 2013.
- [5] Zou Xiaobo, Zhao Jiewen, and Li Yanxiao, "Apple color grading based on organization feature parameters," *Pattern Recognition Letters*, vol. 28, no. 15, pp. 2046–2053, 2007.
- [6] J Blasco, S Cubero, J Gómez-Sanchís, P Mira, and E Moltó, "Development of a machine for the automatic sorting of pomegranate (*punica granatum*) arils based on computer vision," *Journal of Food Engineering*, vol. 90, no. 1, pp. 27–34, 2009.
- [7] M Khojastehnazhand, M Omid, and A Tabatabaeefar, "Development of a lemon sorting system based on color and size," *Afr J Plant Sci*, vol. 4, no. 4, pp. 122–127, 2010.
- [8] V. Leemans, H. Magein, and M.-F. Destain, "On-line fruit grading according to their external quality using machine vision," *Biosystems Engineering*, vol. 83, no. 4, pp. 397 – 404, 2002.
- [9] Vincent Leemans, Hugo Magein, and M-F Destain, "Defect segmentation on jonagoldapples using colour vision and a bayesian classification method," *Computers and Electronics in Agriculture*, vol. 23, no. 1, pp. 43–53, 1999.

- [10] Ali Bulent Koc, "Determination of watermelon volume using ellipsoid approximation and image processing," *Postharvest Biology and Technology*, vol. 45, no. 3, pp. 366 – 371, 2007.
- [11] N. Aleixos, J. Blasco, E. Molto, and F. Navarron, "Assessment of citrus fruit quality using a real-time machine vision system," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 2000, vol. 1, pp. 482–485 vol.1.
- [12] J Blasco, N Aleixos, and E Moltó, "Machine vision system for automatic quality grading of fruit," *Biosystems Engineering*, vol. 85, no. 4, pp. 415 – 423, 2003.
- [13] J. Blasco, N. Aleixos, J. Gómez-Sanchis, and E. Moltó, "Citrus sorting by identification of the most common defects using multispectral computer vision," *Journal of Food Engineering*, vol. 83, no. 3, pp. 384 – 393, 2007.
- [14] J Blasco, N Aleixos, J Gómez-Sanchis, and E Moltó, "Recognition and classification of external skin damage in citrus fruits using multispectral data and morphological features," *Biosystems engineering*, vol. 103, no. 2, pp. 137–145, 2009.
- [15] J. Blasco, N. Aleixos, S. Cubero, J Gómez-Sanchis, and E Moltó, "Automatic sorting of satsuma (citrus unshiu) segments using computer vision and morphological features," *Computers and Electronics in Agriculture*, vol. 66, no. 1, pp. 1 – 8, 2009.
- [16] M. Omid, M. Khojastehnazhand, and A. Tabatabaefar, "Estimating volume and mass of citrus fruits by image processing technique," *Journal of Food Engineering*, vol. 100, no. 2, pp. 315 – 321, 2010.
- [17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [18] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

An Algorithm for In-Place Vision-Based Skewed 1D Barcode Scanning in the Cloud

Vladimir Kulyukin
 Department of Computer Science
 Utah State University
 Logan, UT, USA
 vladimir.kulyukin@usu.edu

Tanwir Zaman
 Department of Computer Science
 Utah State University
 Logan, UT, USA
 tanwir.zaman@aggiemail.usu.edu

Abstract—An algorithm is presented for in-place vision-based skewed 1D barcode scanning that requires no smartphone camera alignment. The algorithm is in-place in that it performs no rotation of input images to align localized barcodes for scanning. The algorithm is cloud-based, because image processing is done in the cloud. The algorithm is implemented in a distributed, cloud-based system. The system's front end is a smartphone application that runs on Android 4.3 or higher. The system's back end is currently deployed on a four node Linux cluster used for image recognition and data storage. The algorithm was evaluated on a set of 506 video recordings of common grocery products. The videos had a 1280 x 720 resolution, an average duration of 15 seconds, and were recorded on an Android Galaxy Nexus smartphone in a local supermarket. The results of the experiments are presented and discussed.

Keywords—computer vision; barcode detection; barcode scanning; mobile computing; skewed barcodes

I. Introduction

According to the World Health Organization (www.who.int), obesity causes such diseases as diabetes, kidney failures, and strokes and predicts that these diseases will be a major cause of death worldwide. Berreby [1] points out that, for the first time in human history, obese people outnumber underfed ones. Such chronic illnesses as diabetes threaten many individuals with numerous complications that include but are not limited to blindness and amputations [2, 3]. The U.S. Academy of Nutrition and Dietetics (www.eatright.org) estimates that approximately twenty six million Americans have diabetes and seven million people in the U.S. are estimated to be aware of their condition. It is estimated that by 2030 the prevalence of diabetes in the world will reach 4.4%, which will equal to approximately 366 million people [3].

While there is no cure for diabetes 1 or 2 as of now, many experts agree that it can be successfully managed. Successful diabetes management has three integral components: healthy diet, blood glucose management, and physical exercise [4]. In this paper, we focus on healthy diet. An important component of a healthy diet is the patients' comprehension and retention of nutritional information and understanding of how different foods and nutritional components affect their bodies. In the U.S. and many other countries, nutritional information is primarily conveyed to consumers through nutritional labels (NLs). Unfortunately, even highly motivated consumers, who

deliberately look for NLs to make healthy food choices, find it difficult to locate them on many products [5].

One way to improve the comprehension and retention of nutritional information by consumers is to use computer vision to scan barcodes in order to retrieve NLs from databases. Unfortunately, a common weakness of many barcode scanners, both open source and commercial, is the camera alignment requirement: the smartphone camera must be aligned with a target barcode to obtain at least one complete scanline for successful barcode recognition [6]. This requirement is acceptable for sighted users but presents a serious accessibility barrier to visually impaired and blind users or to users who may not have good physical command of their hands. Skewed barcode scanning is also beneficial for sighted smartphone users, because it makes barcode scanning faster due to the absence of the camera alignment requirement. Another weakness of the current mobile smartphone scanners is lack of coupling of barcode scanning to comprehensive NL databases from which nutritional information can be retrieved on demand.

To address the problem of skewed barcode scanning, we developed an algorithm for skewed barcode localization on mobile smartphones [7]. In this paper, an algorithm is presented for in-place vision-based skewed barcode scanning that no longer requires the smartphone camera alignment. The algorithm is in-place in that it performs no rotation of input images to align localized barcodes for scanning. The algorithm is cloud-based, because image processing is done in the cloud. The algorithm is implemented in a distributed, cloud-based system. The system's front end is a smartphone application that runs on Android 4.3 or higher. The system's back end is currently deployed on a four node Linux cluster used for image recognition and nutritional data storage.

The front end smartphone sends captured frames to the back end cluster across a wireless data channel (e.g., 3G/4G/Wi-Fi) where barcodes, both skewed and aligned, are recognized. Corresponding NLs are retrieved from a cloud database, where they are stored as HTML documents, and sent across the data channel back to the smartphone where the HTML documents are displayed on the touchscreen. Wikipedia links to important nutrition terms are embedded for better comprehension. Consumers can use standard touch gestures (e.g., zoom in/out, swipe) available on mainstream smartphone platforms to manipulate the NL's surface size. The NL database currently includes approximately 230,000 products compiled from public web sites by a custom crawler.

The remainder of this paper is organized as follows. In Section II, some background information is given on the related work as well as on the research of our laboratory on proactive nutrition management and mobile vision-based nutritional information extraction from product packages. In Section III, we outline the details of our algorithm for in-place skewed barcode scanning in the cloud. In Section IV, we describe our four node Linux cluster for image processing and data storage. Section V presents several experiments with the system and discusses the results. Section VI summarizes our findings, presents our conclusions, and outlines several research venues for the future.

II. Related Work

A. Barcode Localization and Scanning

Much research has been done to on mobile barcode scanning. Tekin and Coughlan [8] describe a vision-based algorithm to guide visually impaired smartphone users to center target barcodes in the camera frame via audio instructions. Wachenfeld et al. [9] present another vision-based algorithm that detects barcodes on a mobile phone via image analysis and pattern recognition methods. A barcode is assumed to be present in the image. The algorithm overcomes typical distortions, such as inhomogeneous illumination, reflections, or blurriness due to camera movement. Unfortunately, the algorithm does not appear to address the localization and scanning of skewed barcodes. Adelman et al. [10] have developed a vision-based algorithm for scanning barcodes on mobile phones. The algorithm relies on the fact that, if multiple scanlines are drawn across the barcode in various arbitrary orientations, one of them might cover the whole length of the barcode and result in a successful barcode scan. This recognition scheme does not appear to handle distorted images, because it is not always possible to obtain the scanlines that cover the entire barcode. Galo and Manduchi [11] present an algorithm for 1D barcode reading in blurred, noisy, and low resolution images. However, the algorithm detects barcodes only if they are slanted by less than 45 degrees. Lin et al. [12] have developed an automatic barcode detection and recognition algorithm for multiple and rotation invariant barcode decoding. The proposed system is implemented and optimized on a DM6437 DSP EVM board, a custom embedded system built specifically for barcode scanning.

A common weakness of many barcode scanners, both open source and commercial (e.g., [13]) is the camera alignment requirement: the smartphone camera must be aligned with a target barcode to obtain at least one complete scanline for successful barcode recognition. This requirement is acceptable for sighted users but presents a serious accessibility barrier to visually impaired shoppers or to shoppers who may not have good dexterity. Skewed barcode scanning is also beneficial for sighted smartphone users, because it may make barcode scanning faster because the camera alignment requirement no longer needs to be satisfied.

B. Proactive Nutrition Management

Many nutritionists and dieticians consider proactive nutrition management to be a key factor in managing diabetes. As more and more individuals start managing their daily activities with

smartphones and other mobile devices, such devices hold great potential to become self-management tools for diabetes and other chronic ailments. Unfortunately, modern nutrition management systems assume that users understand how to collect nutritional data and can be persuaded to perform necessary data collection with emails, SMS's or other digital prompts. Such systems often underperform, because many users find it difficult to integrate nutritional data collection into their daily activities due to lack of time, motivation, or training. Eventually they turn off or ignore digital stimuli [14].

To overcome these challenges, we have begun to develop a Persuasive NUTrition Management System (PNUTS). PNUTS seeks to shift current research and clinical practices in nutrition management toward persuasion, automated nutritional information extraction and processing, and context-sensitive nutrition decision support. The system is based on a nutrition management approach inspired by the Fogg Behavior Model (FBM) [14], which states that motivation alone is insufficient to stimulate target behaviors. Even a motivated user must have both the ability to execute a behavior and a trigger to engage in that behavior at an appropriate place and time. The algorithm presented in this paper is one of the algorithms used by PNUTS for vision-based nutritional information extraction from product packages.

III. Skewed Barcode Scanning

The algorithm for skewed 1D barcode scanning uses our algorithm for skewed barcode localization [7]. The algorithm for skewed barcode localization localizes skewed barcodes in captured frames by computing dominant orientations of gradients (DOG's) of image segments and collecting smaller segments with similar dominant gradient orientations into larger connected components. In Fig. 1, the output of the DOG localization algorithm is shown as a white rectangle around the skewed barcode.

Fig. 2 shows the control flow of our algorithm for skewed barcode scanning. The algorithm takes as input an image captured from the smartphone camera's video stream. This image is given to the DOG algorithm. If the barcode is not localized, another frame is grabbed from the video stream. If the DOG algorithm localizes a barcode, as shown in Fig. 1, the coordinates of the detected region is passed to the line grower component. The line grower component selects the center of the localized region, which is always a rectangle, and starts growing scanlines. For an example of how the line growing component works, consider Fig. 3. In Fig. 3, the horizontal and vertical white lines intersect in the center of the localized region. The skew angle of the localized barcode computed by the DOG algorithm is 120 degrees.

The line that passes the localized region's center at the skew angle detected by the DOG algorithm is referred to as the *skew line*. After the center of the region and the skew angle are determined, the line growing component begins to grow scanlines that are orthogonal to the skew line. In Fig. 3, the skew line is denoted as the black line that passes the region's center at 120 degrees. A scanline is grown on both sides of the skew line. In Fig. 3, the upper half of the scanline is shown as a red arrow and the lower half of the scanline is shown as a blue arrow. Each half is extended until it reaches the portion of the image where the barcode lines are no longer detectable.

A five pixel buffer region is subsequently added after the scanline's end to improve subsequent scanning.



Figure 1. Localization of a skewed barcode

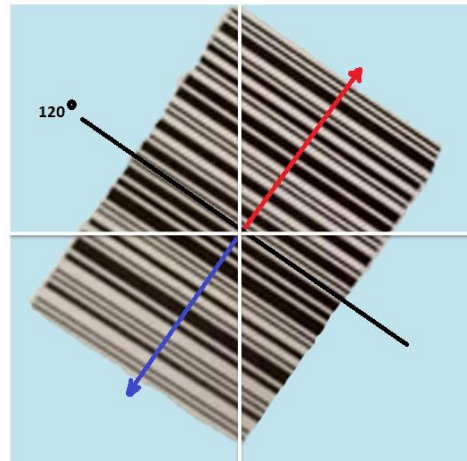


Figure 3. Growth of a scanline orthogonal to the skewed angle of a barcode

The number of scanlines grown on both sides of the skew line is an adjustable parameter. In the current implementation of the algorithm, the value of this parameter is set to 10. The scanlines are arrays of luminosity values for each pixel in their growth path. For each grown scanline, the Line Widths (LW) for the barcode are then computed by finding two points that are on the intensity curve but lie on the opposite sides of the mean intensity. By modelling the curve between these points as a straight line we obtain the intersection point between the intensity curve and the mean intensity.

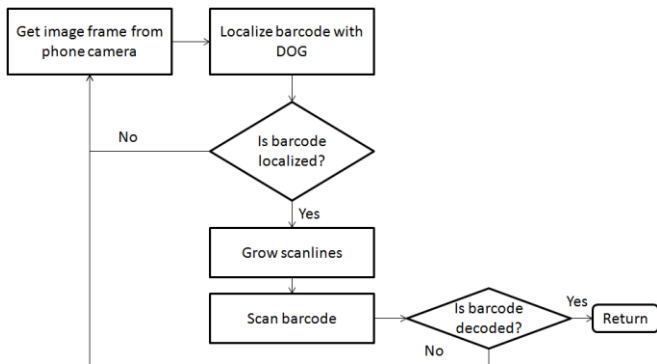


Figure 2. Skewed barcode scanning algorithm

Line Colors (LC) are classified to be either black or white based on whether the pixel intensity is less than or greater than the mean intensity of the scanline. Once the LW and LC are known, each scanline is decoded using the standard EAN decoding scheme. Since UPC is a subset of EAN, this scanning algorithm can decode both EAN and UPC barcodes. The number of scanlines is currently not dynamically adjusted. In other words, ten scanlines are grown and then each of them is passed to the barcode scanner. As soon as a successful scan is obtained, the remaining scanlines, if there are any left, are not scanned. If no scanlines result in a successful scan, the control returns back to capturing frames from the video stream.



Figure 4. Sequence of images that demonstrates how the algorithm works

Fig. 4 shows a sequence of images that gives a visual demonstration of how the algorithm processes a captured frame. The top image in Fig. 4 is a frame captured from the

smartphone camera's video stream. The second image in Fig. 4 shows the result of the clustering stage of the DOG algorithm that clusters small subimages with similar dominant gradient orientations. The third image in Fig. 4 shows a white rectangle that denotes the localized barcode. The fourth image in Fig. 4 shows the ten scanlines, one of which results in a successful skewed barcode scan.

IV. Linux Cluster for Image Processing and Data Storage

We built a Linux cluster out of four Dell computers for cloud-based computer vision and data storage. Each computer has an Intel Core i5-650 3.2 GHz dual-core processor that supports 64-bit computing. The processors have 3MB of cache memory. The machines are equipped with 6GB DDR3 SDRAM and have Intel integrated GMA 4500 Dynamic Video Memory Technology 5.0. All machines have 320 GB of hard disk space. Ubuntu 12.04 LTS was installed on each machine.

We used JBoss (<http://www.jboss.org>) to build and configure the cluster and the Apache mod_cluster module (http://www.jboss.org/mod_cluster) to configure the cluster for load balancing. Our cluster has one master node and three slaves. The master node is the domain controller. The master node also runs mod_cluster and httpd. All four machines are part of a local area network and have hi-speed Internet connectivity. We have installed JDK 7 in each node.

The JBoss Application Server (JBoss AS) is a free open-source Java EE-based application server. In addition to providing a full implementation of a Java application server, it also implements the Java EE part of Java. The JBoss AS is maintained by jboss.org, a community that provides free support for the server. JBoss is licensed under the GNU Lesser General Public License (LGPL).

The Apache mod_cluster module is an httpd-based load balancer. The module is implemented with httpd as a set of modules for httpd with mod_proxy enabled. This module uses a communication channel to send requests from httpd to a set of designated application server nodes. An additional communication channel is established between the server nodes and httpd. The nodes use the additional channel to transmit server-side load balance factors and lifecycle events back to httpd via a custom set of HTTP methods collectively referred to as the Mod-Cluster Management Protocol (MCMP).

The mod_cluster module provides dynamic configuration of httpd workers. The proxy's configuration is on the application servers. The application server sends lifecycle events to the proxies, which enables the proxies to auto-configure themselves. The mod_cluster module provides accurate load metrics, because the load balance factors are calculated by the application servers, not the proxies.

All nodes in our cluster run JBoss AS 7. Jboss AS 7.1.1 is the version of the application server installed on the cluster. Apache httpd runs on the master with the mod_cluster-1.2.0 module enabled. The Jboss AS 7.1.1 on the master and the slaves are discovered by httpd.

A Java servlet for image recognition is deployed on the master node as a web archive file. The servlet's URL is hardcoded in every front end smartphone. The servlet receives images uploaded with http post requests, recognizes barcodes, and sends an HTML response back to front end smartphones.

No data caching is currently done on the servlet or the front end smartphones.

V. Experiments

The skewed barcode scanning experiments were conducted on a set of 506 video recordings of common grocery products that we have made publicly available [15]. The videos have a 1280x720 resolution, an average duration of 15 seconds and were recorded on an Android 4.2.2 Galaxy Nexus smartphone in a supermarket in Logan, UT. All videos were taken by an operator who held a grocery product in one hand and a smartphone in the other. The videos covered four different categories of products: bags, boxes, bottles, and cans.

Colored RGB frames were extracted from each video at the rate of one frame per second and grouped together into different categories of products. Each frame was automatically classified as blurred or sharp by the blur detection scheme using Haar wavelet transforms [16, 17] that we implemented in Python. Each frame was manually classified as having a barcode or not.

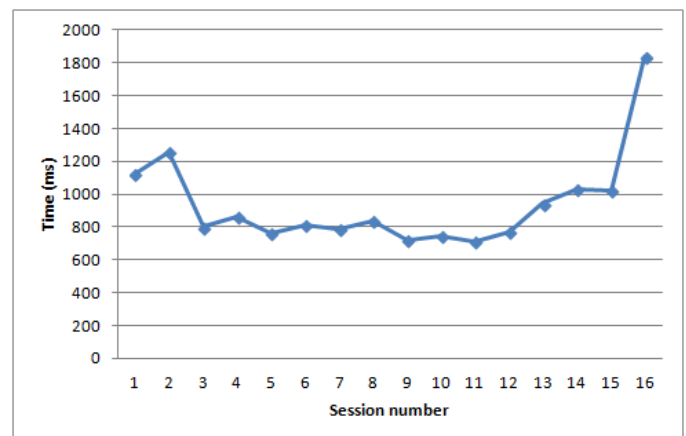
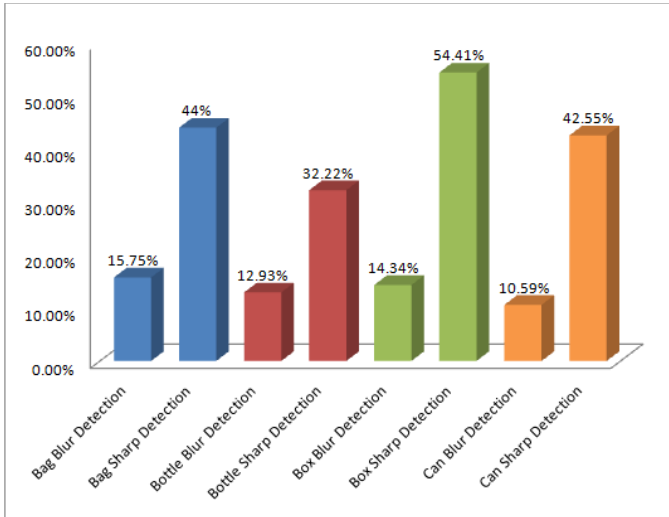


Figure 5. Average request-response times in milliseconds

After all classifications were completed (blurred vs. sharp, barcode vs. no barcode), the classified frames were stored in the smartphone's sdcard. The evaluation procedure was implemented as follows. A started Android service would take one frame at a time and uploaded it to the Linux cluster via an http POST request over the USU Wi-Fi network. The network has a download speed of 72.31 Mbps and an upload speed of 29.64 Mbps.

Each captured frame was processed on the cluster as follows. The DOG localization algorithm [7] was executed and, if a barcode was successfully localized, the barcode was scanned within the localized region in place with ten scanlines. The detection result was sent back to the smartphone and recorded on the smartphone's sdcard. The average request-response time for each session was calculated. Fig. 5 gives the graph of the node cluster's request-response times. The lowest average was 712 milliseconds; the highest average was 1813 milliseconds.



VI. Results

Images that contained barcodes for all four product categories had no false positives. As shown in Fig. 6, for each product category, the sharp images had a significantly better true positive percentage than the blurred images. A comparison of the bar charts in Fig. 6 reveals that the true positive percentage of the sharp images is more than double that of the blurry ones. Images without any barcode for all categories produced 100% accurate results with all true negatives, irrespective of the blurriness. In other words, the algorithm is highly specific in that it does not detect barcodes in images that do not contain them.

Another observation on Fig. 6 is that the algorithm showed its best performance on boxes. The algorithm's performance on bags, bottles, and cans was worse because of crumpled, curved, or shiny surfaces. These surfaces caused many light reflections and hindered performance of the skewed barcode localization and scanning. The percentages of the skewed barcode localization and scanning were better on boxes due to smoother surfaces. Quite expectedly, the sharpness of a frame also makes a positive difference in that the algorithm performed much better on sharp images in each product category. Specifically, on sharp images, the algorithm performed best on boxes with a true positive percentage of 54%, followed by bags at 44%, bottles at 32%, and cans at 30%.

As Fig. 7 shows, that the average scan time was lowest for boxes and largest for cans. This finding is in line with the results in Fig. 6. The average scan times for cans and bags were significantly longer than for boxes. The average scan time for bottles was longer than for boxes but shorter than for cans and bags. However, the explanation is not limited to crumpled, curved, and shiny surfaces. Another reason for the slower scan times on individual products in each product category is the availability of Internet connectivity at various locations in the supermarket. During our scan experiments in the supermarket, we noticed that at some areas of the supermarket the Internet connection did not exist, which caused delays in barcode scanning. For several products, a 10- or 15-step change in location within a supermarket resulted in a successful barcode scan.

VII. Discussion

The skewed barcode scanning algorithm presented in this paper targets medium- to high-end mobile devices with single or quad-core ARM systems. Since cameras on these devices capture several frames per second, the algorithm is designed to minimize false positives rather than maximize true ones, because, at such frequent frame capture rates, it is far more important to minimize the processing time per frame. In other words, the algorithm is highly specific, where specificity is the percentage of true negative matches out of all possible negative matches.

Figures 8, 9, 10, and 11 demonstrate the specificity of the algorithm on all four categories of products. In all product categories, the true negative and false positive percentages are 0, which means that the algorithm accurately does not recognize barcodes in images that do not contain them. In all categories, the false negative percentages are relatively high. This is done by design. The algorithm is designed to be

Figure 6. Skewed barcode scanning in blurred and sharp images

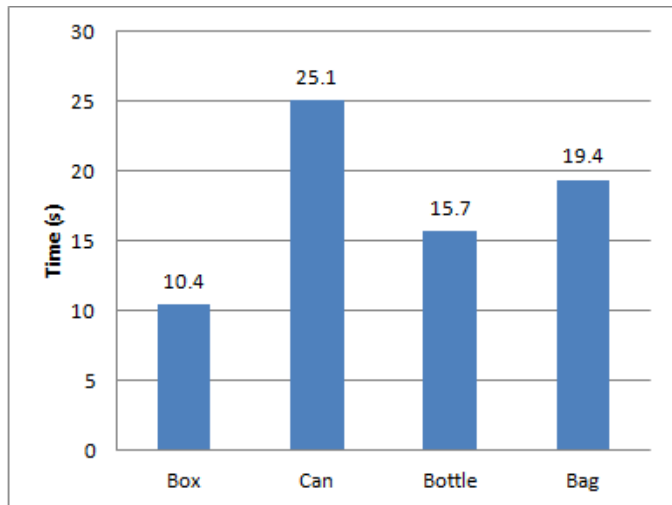


Figure 7. Average barcode scan times for each product category in a local supermarket in seconds

We also conducted skewed barcode scanning experiments in a local supermarket. A sighted user was given a Galaxy Nexus 4 smartphone with an AT&T 4G connection. Our front end application was installed on the smartphone. The user was asked to scan ten products of his choice in each of the four categories: box, can, bottle, and bag. The user was told that he can choose which products to scan. A research assistant accompanied the user and recorded the scan times for each product. Each scan time started from the moment the user began scanning and ended when the response was received from the server. Fig. 7 denotes the average times in seconds for each category.

conservative in that it rejects the frames on the slightest chance that it does not contain any barcode. While this increases false negatives, it keeps both true negatives and false positives close to zero.

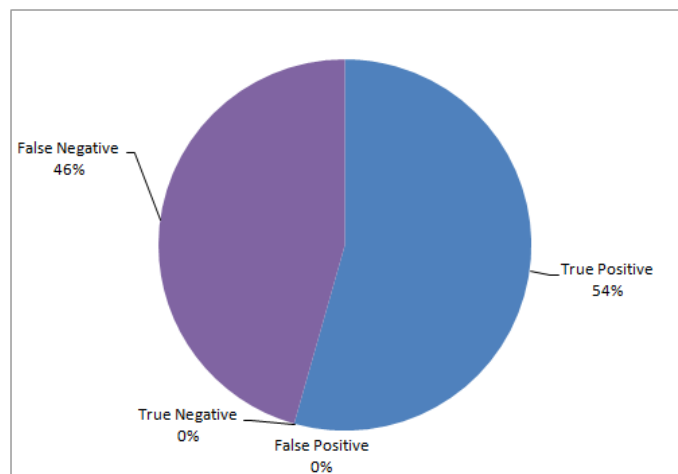


Figure 8. Performance on sharp box images

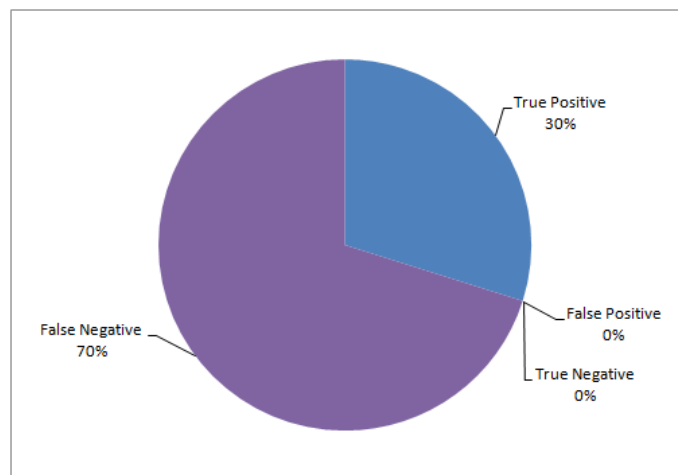


Figure 9. Performance on sharp can images

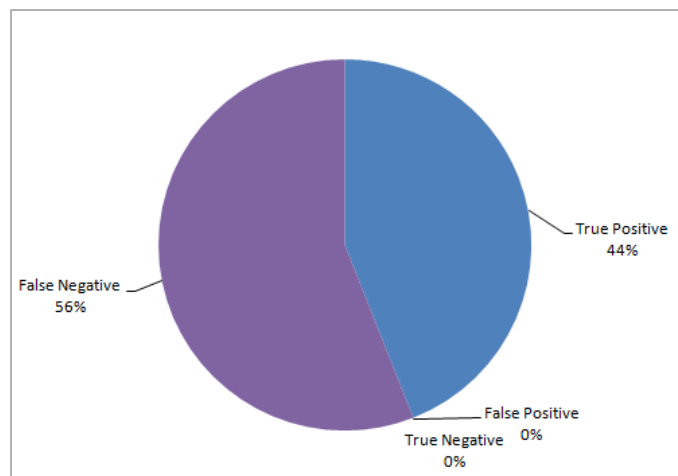


Figure 10. Performance on sharp bag images

A limitation of the current implementation of this algorithm which was discovered during our field experiments

in a supermarket is that there is no run-time checking for the availability of Internet connectivity. If such checking is implemented, the user can be quickly notified that there is no Internet connectivity and the frame grabbing from the video stream can be temporarily halted.

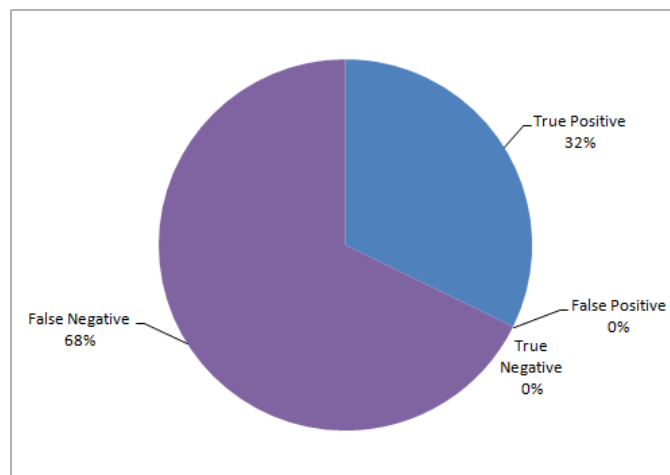


Figure 11. Performance on sharp bottle images

Another limitation of the current implementation is that it does not compute the blurriness of the captured frame before sending it to the cluster for barcode localization and detection. As shown in Fig. 6, the scanning results are substantially higher on sharp images than on blurred images. This limitation points to a potential improvement that we plan to implement in the future. When a frame is captured, its blurriness coefficient can be computed on the smartphone and, if it is high, the frame should not even be sent to the cluster. This improvement will also reduce the load on the cluster and may increase response times.

Another approach to handling blurred inputs is to improve camera focus and stability, both of which are outside the scope of this algorithm, because it is, technically speaking, a hardware problem. It is likely to work better in later models of smartphones. The current implementation on the Android platform attempts to force focus at the image center but this ability to request camera focus is not present in older Android versions. Over time, as device cameras improve and more devices run newer versions of Android, this limitation will have less impact on recall but it will never be fixed entirely.

Finally, we would like to implement a tighter integration of the current system with proactive nutrition management. Specifically, we plan to integrate skewed barcode scanning with a wireless glucometer so that nutrition intake recording can be coupled with glucometer readings. This will allow users to actively monitor the impact that various foods have on the level of glucose in their blood.

References

[1] Anding, R. *Nutrition Made Clear*. The Great Courses, Chantilly, VA, 2009.
 [2] D. Berreby. "The obesity era." *The Aeon Magazine*. June 19, 2013.

- [3] Rubin, A. L. *Diabetes for Dummies*. 3rd Edition, Wiley, Publishing, Inc. Hoboken, New Jersey, 2008.
- [4] Eirik Årsand, E., Tatara, N., Østengen, G. and Gunnar Hartvigsen, G. "Mobile phone-based self-management tools for type 2 diabetes: the few touch application." *Journal of Diabetes Science and Technology*, Vol. 4, Issue 2, March 2010.
- [5] Graham, D. J., Orquin, J. L., and Visshers, V. H. M. "Eye tracking and nutritional label use: a review of the literature and recommendations for label enhancement." *Food Policy*, vol. 32, pp. 378-382, 2012.
- [6] Kulyukin, V., Kutiyawala, A., and Zaman, T. "Eyes-Free Barcode Detection on Smartphones with Niblack's Binarization and Support Vector Machines." In Proceedings of the 16-th International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV 2012), Vol. I, pp. 284-290, CSREA Press, July 16-19, 2012, Las Vegas, Nevada, USA. ISBN: 1-60132-223-2, 1-60132-224-0.
- [7] Kulyukin, V. and Zaman T. "Vision-Based Localization of Skewed UPC Barcodes on Smartphones." In Proceedings of the International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV 2013), pp. 344-350, pp. 314-320, ISBN 1-60132-252-6, CSREA Press, Las Vegas, NV, USA.
- [8] Tekin, E. and Coughlan, J. "A mobile phone application enabling visually impaired users to find and read product barcodes." In Proceedings of the 12th International Conference on Computers Helping People with Special needs (ICCHP'10), Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 290-295, 2010.
- [9] Wachenfeld, S., Terlunen, S., and Jiang, X. "Robust recognition of 1-D barcodes using camera phones." In Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008), pp.1-4, Dec. 8-11, 2008. ISSN: 1051-4651, IEEE.
- [10] Adelman R., Langheinrich M., Floerkemeier, C. A Toolkit for BarCode Recognition and Resolving on Camera Phones - Jump Starting the Internet of Things. Workshop on Mobile and Embedded Information Systems (MEIS'06) at Informatik 2006, Dresden, Germany, Oct 2006.
- [11] Gallo, O.; Manduchi, R., "Reading 1D barcodes with mobile phones using deformable templates." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.33, no.9 ,pp.1834-1843,Sept.2011. doi:10.1109/TPAMI.2010.229.
- [12] Lin, D. T., Lin, M.C., and Huang, K.Y. 2011. "Real-time automatic recognition of omnidirectional multiple barcodes and DSP implementation." *Mach. Vision Appl.* Vol 22, num. 2, pp. 409-419, March 2011. doi:10.1007/s00138-010-0299-3.
- [13] Zxing, <http://code.google.com/p/zxing/>.
- [14] Fog, B.J. "A behavior model for persuasive design." In Proc. 4th International Conference on Persuasive Technology, Article 40, ACM, New York, USA, 2009.
- [15] Mobile Supermarket Barcode Videos. <https://www.dropbox.com/sh/q6u70wcg1luxwdh/LPtUBdwdY1>.
- [16] Tong, H., Li, M., Zhang, H., and Zhang, C. "Blur detection for digital images using wavelet transform," In Proceedings of the IEEE International Conference on Multimedia and Expo, Vol.1, pp. 27-30, June 2004. doi: 10.1109/ICME.2004.1394114.
- [17] Niavegelt, Y. *Wavelets Made Easy*. Birkhäuser, Boston, 1999.

SESSION
IMAGING SCIENCE AND MEDICAL
APPLICATIONS

Chair(s)

TBA

Sickle Anemia and Distorted Blood Cells Detection Using Hough Transform Based on Neural Network and Decision Tree

Hany A. Elsalamony¹

¹Mathematics Department, Faculty of Science, Helwan University, Cairo, Egypt.

h_salamony@yahoo.com

Abstract - Sickle-cell anemia is one of the most important common types of anemia disease. This paper presents proposed algorithm in two parts, one is the construct an algorithm can detecting and counting RBCs (benign or distorted) in a microscopic colored image; even if they are hidden or overlapped. Second part is checking and analysing the constructed data of RBCs by applying the most common important techniques in data mining; neural network and decision tree. The experimental results are demonstrated high accuracy, and success using these two models in predicting for all the benign or distorted cells. This algorithm has achieved the highest segmentation by about 99.98% of all input cells, which is contributed to improve the diagnosis of Sickle anemia. The Neural Network has agreed with the detection by the proposed algorithm in prediction outcome by about 96.9%, whereas the Classification and Regression (C&R) tree has achieved 92.9%.

Keywords: Sickle Anemia; Image watershed segmentation; Red Blood Cells' detection and counting; C&R tree; Neural Network.

1 Introduction

Nowadays, the analysis of the blood cells' microscopic image is very impressive diagnostic tool for many diseases. Actually, human blood is a complex combination of plasma, red blood cells (RBCs), white blood cells (WBCs), and platelets. Plasma is the fluid component, which is contained melted salts and proteins. RBCs make up about 40% of blood volume. WBCs are less, but greater in size than RBCs. The platelet cells are similar particles, which are smaller than WBCs and RBCs [16].

Anemia is occurred when the blood has a lower than the normal number of red blood cells (RBCs) or they have not been enough hemoglobin. RBCs have been made inside the larger bones of the body in the spongy marrow. Bone marrow is always making new red blood cells to replace old ones. Normal RBCs die after they living 120 days in the bloodstream. Their job is carrying oxygen and removing carbon dioxide (a waste product) from the body. In addition, RBCs are disc-shaped and moving easily through blood vessels. They contain an iron-rich protein called hemoglobin. This protein transmits oxygen from

the lungs to the rest of the body [16]. In Sickle-cell anemia, the body makes sickle-shaped red blood cells in a serious disorder. Sickle cells contain abnormal hemoglobin called Sickle hemoglobin or hemoglobin S, which it helps the cells to develop a sickle, or crescent, shape. In fact, Sickle cells are very dangerous because rigidity and sticky of them, that are tended to block blood flow in the blood vessels of the limbs and organs. The blocked blood flow can cause pain, organ damage, and can raise the probability of infection. Moreover, the abnormal Sickle cells usually die after only about 10 to 20 days. The bone marrow cannot make new red blood cells fast enough to replace the dying ones [16]. Figure.1 illustrates the danger of Sickle anemia and kinds of it.

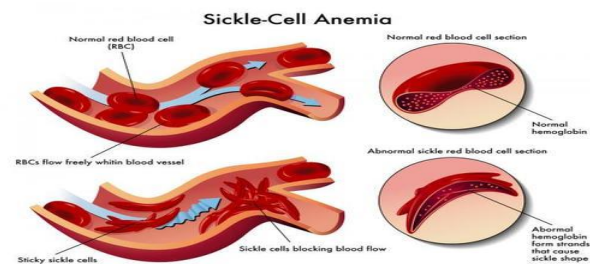


Figure.1 Kinds of Sickle's anemia and RBCs

In addition, the Sickle-cell anemia is most common in people whose families derived from Mediterranean countries, Africa, south or Central America, especially Panama, Caribbean islands, Saudi Arabia, and India. The United States estimated that the infection from 70,000 to 100,000 people, mainly African Americans. The main discovery of this disease is depending on blood test analysis that can detect Sickle cells.

Recently, the microscopic image analysis helps as an impressive diagnostic tool for the infected blood cells' detection [16]. The Hough transform is the most important technique in image analysis and segmentation to help in that way. It is depending on extracting features related through the segmentation process of microscopic image. Hough transform that is generally used today was invented by Richard Duda and Peter Hart in 1972, who called it a "generalized Hough transform" after the related 1962 patent of Paul Hough [8], [12].

On the other hand, Data mining techniques have gained popularity and useful for illustrative and predictive applications of image analysis. Two techniques are applied to the resulted data on blood cells' detection. Neural network (NN) is one of these techniques; it has been applied successfully in the identification and control of dynamic systems. Previously, the progress of for NN began by David Rumelhart in 1986. He presented the back propagation that distributed pattern recognition errors and using many layers throughout the network.

Another one of data mining applied techniques is the classification and regression tree (C&R) as the most common and powerful technique for the classification and prediction in the decision tree (DT). It can generate understandable rules, and to handle both continuous and categorical variables [7]. Historically, the seminal book by Breiman et al. (1993) provided an introduction to decision trees that is still considered the standard resource on the topic. The C&R tree algorithm popularized by Breiman, Friedman, Olshen, and Stone in 1984, and by Ripley in 1996 [3]. Two reasons behind the popularity of decision tree techniques are (1) the procedures are relatively straightforward to the understand and explain, and (2) the procedures address a number of data complexities, such as nonlinearly and interactions, that commonly occur in real data [7].

The work of this paper is divided into two parts; one is applying a proposed algorithm to detect the benign and distorted blood cells (Sickle anemia) and counting them depending on segmentation of their shapes using Circular Hough Transform, watershed, and morphological tools. The second part is introducing an algorithm for checking and analysis of the resulted cells' data variables (Areas, Convex Area, Perimeter, Eccentricity) by applying the common most important techniques in data mining; neural network and decision tree (classification and regression tree) to get the right decision for diagnoses [2].

The rest of the paper is organized as follows; Section (2) focuses on the related work, and the definition and features of Hough transform is presented in section (3). Section (4) overview of the neural network. Section (5) is presented an overview of classification and regression of decision trees. The proposed algorithm is discussed in section (6). The experimental results show the effectiveness of each model in section (7). The conclusion has been presented in section (8).

2 Related work

In recent years, research on blood cells' recognition and diagnosis of diseases has grown rapidly. In May 2013, K. Thirusittampalam, M. J. Hossain, O. Ghita, and P. F. Whelan, developed a novel-tracking algorithm that can extract the cell motility indicators and determined the cellular division (mitosis) events in large time-lapse phase-contrast image sequences. Their process of automatic unsupervised cell tracking carried out in a sequential manner, where the inter

frame cell's association is achieved by assessing the variation in the local cellular structures in consecutive frames from the image sequence. The experimental results indicated that their algorithm achieved 86.10% overall tracking accuracy and 90.12% mitosis detection accuracy [9].

Another proposal introduced by H. A. Khan and G. M. Maruf in May 2013, that presented an algorithm for cell segmentation and counting by detection of cells' centroids in microscopic images. The method is specifically designed for counting circular cells with a high probability of occlusion. The experimental results showed an accuracy of 92% of cell counting even at a very high 60% overlap probability [5].

An algorithm presented by M. C. Mushabe, R. Dendere and T. S. Douglas in July 2013, that they were identified and counted red blood cells (RBCs) as well as parasites in order to perform a parasitemia calculation. Morphological operations and histogram-based threshold were used to detect the red cells. They used boundary curvature calculations and Delaunay triangulation to split overlapped red cells. The parasites are classified by Bayesian classifier with their RGB pixel values as features. The results showed 98.5% sensitivity and 97.2% specificity for detecting infected red blood cells [10]. The next section introduces an overview on Hough transform.

3 Hough Transform

The Hough Transform is a popular feature extraction technique that converts an image from its Cartesian to its Polar coordinates. Any point within the image space is represented by a sinusoidal curve in the Hough space. In addition, two points to a line segment generate two curves, which are overlaid at a location corresponds with a line through the image space. Even though this model form is very easy it is deeply complicated for the case of complex shapes due to noise and shape imperfection, also the problem of finding slopes of vertical lines. Circular Hough Transforms (CHT) solved this problem by putting a transformation of the centroid of the shape in the x-y plane to the parameter space [6].

However, there are three essential steps, which are common to all CHT: first one is an Accumulator Array Computation, which is working as that foreground pixel of a high gradient are chosen as being candidate pixels and are allowed to 'votes' in the accumulator array. Center estimation is the second step; the circle centers are expected by detecting the peaks in the accumulator array by voting of candidate pixels that are belonging to an image circle tend to accumulate in the accumulator array box corresponding to the circle's center.

Figure.2 shows an example of the candidate pixels (solid dots) falling on an actual circle (solid circle), and their voting patterns (dashed circles) which coincide with the center of the substantial circle. The third step in CHT is radius estimation; that is if the same accumulator array has used for more than one radius value, as is commonly done in CHT algorithms, the radii of detected circles have estimated as a separate step [8].

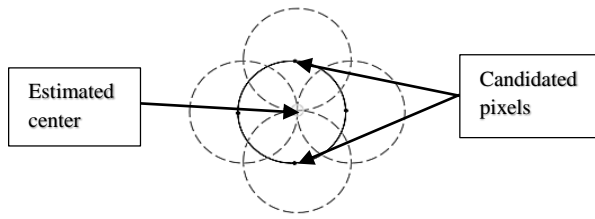


Figure.2 Circle center estimated.

The radius has estimated clearly by using radial histograms; however in Phase-Coding, the radius can be estimated by simply decoding phase information from the estimated center located within accumulator array [8]. This paper has used CHT to detect and count RBCs; even if they are hidden or overlapped. Watershed and morphological functions are used for enhancing and separating overlapped cells during the segmentation process.

4 Overview of Neural Network

At the present time, Neural Network (NN) with back propagation is the most popular artificial neural network construction. It is known as a powerful function approximation for prediction and classification problems. Historically, NN is a mutually dependent group of artificial neurons that uses a mathematical model for information processing with a connected approach to computation by Freeman in 1991 [4], [7]. The NN structure is organized into layers of input, output neurons, and hidden layers. The activation function may range from a simple threshold function, or a sigmoid, hyperbolic tangent, or radial basis function [1].

$$y_i = f(\sum w_{ij}x_i) \quad (1)$$

The back propagation is a common training technique for NN. This training process is to perform a particular function by adjusting the values of the connections (weights) between elements [2], [17]. Actually, three important issues in NN need to be addressed; selection of data samples for network training, selection of an appropriate and efficient training algorithm, and determination of network size [14], [15]. Moreover, NN has many advantages, such as the good learning ability, less memory demand, suitable generalization, fast real-time operating, simple and convenient to utilize, suited to analyze complex patterns, and so on. On the other hand, there are some disadvantages like that it requires high-quality data; variables must be carefully selected a priori, the risk of over-fitting, and requires a definition of architecture [4].

5 Overview of Classification and Regression Tree

Classification and regression trees (C&R) are the most common and popular non-parametric decision tree learning technique. In this paper, a regression tree only uses for numeric data values. C&R builds a binary tree by splitting the records at each node according to a function of a single input variable.

The measure that used to evaluate a potential splitter is diversity. The best splitter is the one that decreases the diversity of the record sets by the great one. This method uses recursive partitioning to split the training records into segments with similar output variable values [7]. Moreover, the impurity that used in each node defined in the tree by two measures; entropy, as in Equation (2), and Gini (chosen in this paper).

$$Entropy(t) = -\sum_j p(j|t)\log p(j|t) \quad (2)$$

The Gini index generalizes the variance impurity, that variance is of distribution related to the two classes. However, the Gini index, as in Equation (3), can also be useful as the expected error rate if the class label is randomly chosen from the class distribution at the node. This impurity measure has been slightly stronger at equal probabilities (for two classes) than the entropy measure. Gini holds some advantage for an optimization of the impurity metric at the nodes [11].

$$g(t) = 1 - \sum_j p^2(j|t) \quad (3)$$

When the cases in a node are evenly distributed across the categories; the Gini index takes its maximum value of $1-(1/k)$, where k is the number of categories for the target attribute. Furthermore, for all cases in the node, which are belonged to the same category the *Gini* index equals zero. The proposed algorithm is displaying in the next section.

6 The proposed algorithm

The goal of this paper is to detect benign and distorted blood cells in a colored microscopic image and distinguishing between them. This work can help doctors, physicians, chemists... who cares about blood cell detection, analysis, and determination of diseases. The proposed algorithm is divided into two steps, as shown in Figure.3. One is applying CHT with morphological functions on bright and dark of intensity cells to detect and count benign and distorted blood cells. The second step; NN and C&R tree are applied to test and check the performance of the proposed algorithm for diagnosing and deciding that the patient has Sickle anemia or not and which it is more effective than another. The classification and prediction are depended on the detected cells' data variables (Areas, Convex Area, Perimeter, Eccentricity) to reduce the errors in detection operation and ensure about final diagnosis for Sickle Anemia existence. In the first part of CHT many operations have been carried out through it:

- Cell polarity; indicates whether the circulating blood cells are brighter or darker than the background.
- Computation method (Two-stage) is used to calculate the accumulator array of CHT. It is based on computing radial histograms; radii are clearly applying the estimated cell centers along with the image information [13].
- Sensitivity factor is the sensor of the accumulator array in CHT. The detection is including weak and partially hidden or overlapped cells; however higher values of the sensitivity increasing the risk of false detection.

- Edge gradient threshold; cells have generally a darker interior (nuclei) and surrounded by an outlying bright halo. The edge gradient threshold is very useful for determining edge pixels in these cases of image, both weak and strong blood cells based on their contrast are detected well by setting a lower value in the threshold. It detects fewer cells of weak edges by increasing the value of the threshold [1].

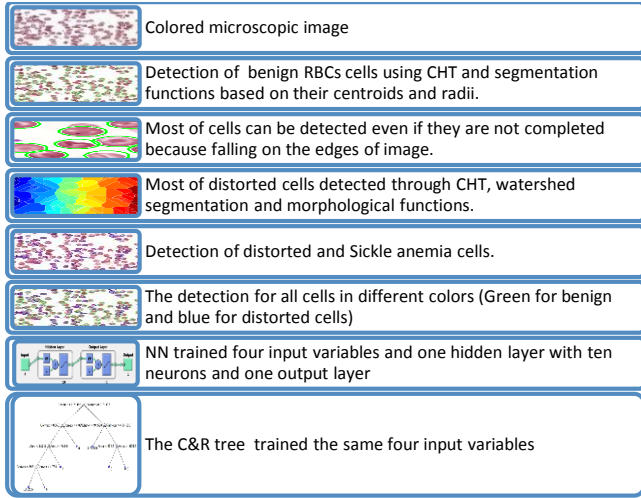


Figure.3 the main points of the proposed algorithm

The final stage is the segmentation process; is displaying of input image with all the contoured benign (green) and distorted (blue) RBCs. Through this operation; the area, convex area, perimeter, eccentricities for each cell are measured, also the good and distorted cells can be counted and diagnosed with Sickle anemia. These measures are used as input variables to train NN and C&R tree while the output (target) is measured based on the solidity (S), which is a division cell area over its convex area for all cells, as in the Equation (4):

$$S = \frac{area_c}{(convex\ area_c)} \quad (4)$$

$$T_c = \begin{cases} 1 & \text{if } 0.95 < S_c \leq 1 \\ 0 & \text{if } 0 \leq S_c \leq 0.95 \end{cases} \quad (5)$$

where $area_c$ is the area in each cell, and $convex\ area_c$ represents its convex area for all detected cells (benign and distorted). The target T_c have been two values 1 and 0 based on the solution of Equation (4); such that if any of solidity values S_c is greater than 0.95 up to 1 (perfect cell) T_c takes the value 1 with the decision benign. On the other hand, if T_c takes the value 0, then S_c value has less than or equal 0.95 according to Equation (4) with the decision that the cell is distorted and may be sickled.

The back propagation neural network has been trained and tested four variables as an input layer and ten neurons in the hidden layer whereas one neuron in the output layer. Moreover, three kinds of samples are applied training, validation, and testing samples. The training samples are presented in the network during the training process by 80% from all samples in the input variables, and the network is modified according to

its error. Accordingly, only 10% are used for the validation samples to measure network generalization and to pause training when generalization stops improving, and the remaining 10% of all samples of cells are introduced to be a testing sample that have no effect on training and so provide an independent measure of network performance during and after training. In addition, the mean square error (MSE) is applied, that it is defined as the average squared difference between outputs and targets, whereas lower values are better, Zero means no error. Table 1 illustrates the description of the formed cells' data, which is automatically computed for all cells (benign and distorted).

TABLE 1 BLOOD CELLS' VARIABLES

#	Variables	Type	Domain
1	Areas	Ranged	163:1376
2	Convex area	Ranged	187:1781
3	Perimeter	Ranged	58.97056:218.2082
4	Eccentricity	Ranged	0:0.965444268
5	Target	Binary	(1 for benign, 0 for distorted cells)

On the other way, the recursive binary C&R tree has been applied to the term of regression because all the variables containing numeric values, as in Table 1. In this Table, all variables have ranged in different types whereas only the output (target) variable has binary values such as 1 values for benign cells and 0 values for all distorted cells. The binary tree has divided into two branches based on the Gini index recursively trained with maximum tree depth five levels and stopping when a minimum records in parent branch 2% and minimum records in child branch 1%.

Finally, the performance of each classification model is evaluated using three statistical measures: classification accuracy, sensitivity and specificity. These measures are defined as true positive (TP), true negative (TN), false positive (FP) and false negative (FN). A true positive decision occurs when the positive prediction of the classifier coincided with a positive prediction of the previously segmentation. A true negative decision occurs when both the classifier and the segmentation suggests the absence of a positive prediction. False positive occurs when the system labels benign cell (positive prediction) as a malignant or distorted one. Finally, false negative occurs when the system labels a negative (malignant) cell as positive. Moreover, the classification accuracy is defined as the ratio of the number of correctly classified cells and is equal to the sum of TP and TN divided by the total number of RBCs (N) [9].

$$Accuracy = \frac{TP+TN}{N} \quad (6)$$

The sensitivity refers to the rate of correctly classified positive and is equal to TP divided by the sum of TP and FN .

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

Specificity refers to the rate of correctly classified negative and is equal to the ratio of TN to the sum of TN and FP [9].

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (8)$$

The next section is displaying the experimental results of the proposed algorithm in details.

7 The experimental results

As mentioned before, the experimental results are displayed in two parts; one is concerned in the RBCs detection and segmentation to distinguish between benign and distorted cells (caused by Sickle Anemia) and counted automatically. The work of second part is the checking the previous detection process exported from part one by predicting and categorizing the segmentation results using the most famous two classification models in data mining NN and C&R tree [2]. Accordingly, the proposed algorithm has provided the ratio of errors during the detection process and performance of success using CHT in this case of RBCs detection but not effective.

In part one; the detection process has been started by importing and reading the microscopic colored RBCs image. This detection for all cells (benign or distorted) is done using CHT, watershed process, morphological techniques for enhancing the detection process. In the same way, CHT is applied under the conditions of cell polarity to determine all dark and bright cells according to their intensity. Then, two-stage technique that used to compute the accumulator array of CHT; the sensitivity of this accumulator array of the proposed algorithm is 0.97 for brightness and 0.90 for dark cells, and Edge gradient threshold 0.2 that detect fewer cells with weak edges. Actually, these conditions help CHT to detect most of the benign RBCs (near to the circle in shape), those positioned singular, overlapped, and even those are attached to other distorted cells. Figure.4 shows the original image of RBCs in (a), and the proposed algorithm in details is illustrated from (b) to (f).

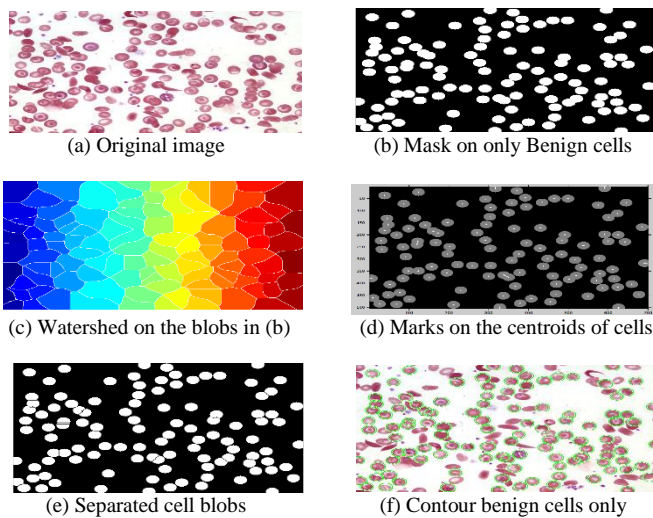


Figure.4 (a) The original image, (b) The healthy cells masked, (c) Watershed for each blob region, (d) Makes a small mark on every cell's centroid, (e) The cell's blobs separated, and (f) The final detected benign cells.

In (b); a mask of balls constructed using centroids and radii for each cell, which have determined before; in this case and after the masking process, many overlapping benign cells appeared as a one big region according to the shape and size of normal cells. For that reason, the watershed process in (c) is applied to get the optimum separation in this abnormal shape of cell. In fact, to get the optimal separation of overlapped cells their centroids are used to put marks on them as in (d). Now the separation has been ready regarding these marks for each cell blob as in (e). Finally, by applying all the previous steps; the only benign cells are contoured by the green line, extracted to count, and distinguished from the other distorted cells, as well shown in the second step of part one. Through this detection operation of benign cells; the number of these cells is equal to 109 benign cells out of 180 total number of all detected cells (benign and distorted) in this image. In fact, the remaining number of cells (71) may be considered as detection errors, distorted cells (Sickle Anemia), platelets or even WBCs. This image has not any WBCs; but if they exist in other images, the proposed algorithm can easily detect and count them. On the other hand, platelets have been neglected because the algorithm concentration only on RBCs and distorted Sickle Anemia disease.

Additionally, the algorithm has been determined 177 cells as a total of detecting blood cells by 99.98% success ratio according to the image in Figure.4 (a). Therefore, the next step is trying to know that how many cells out of 71 are Sickle cells or initiated to be it. Firstly, all these 71 strange shapes (crescent, elliptic, platelets, and unknown) are discovered and displayed using the same previously steps of benign cell detection. Figure.5 illustrates last two steps of the proposed algorithm, which is applied to the distorted cells. In (a) a colored segmented image for all unknown distorted shapes with a noise like platelets ... etc. The Sickle cells or the cells that initiated to be Sickle are detected without any noise in (b). On the other hand, the only deformed cells are counted 57 (Sickle or initiated to be) out of 71. The last detection of distorted cells is contoured by blue line shown in (c). In the (d) of Figure.5, the final detection and tracking of all cells (benign by green color and distorted with blue color) has been completed.

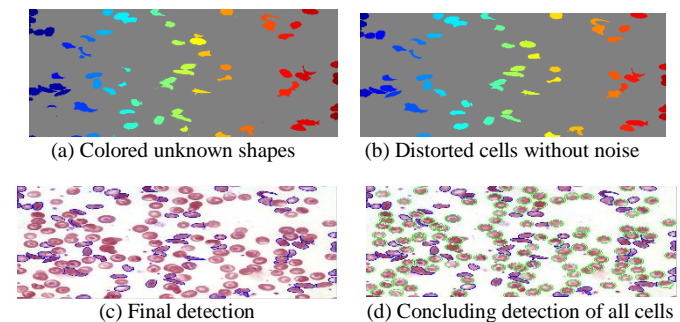


Figure.5 (a) All unknown shapes, (b) the distorted cells, (c) the final detection of only distorted cells, (d) the final detection of all cells.

After the end of this part of detection and segmentation, the second part of the proposed algorithm is began that concerning prediction and analysis of resulted cells' detection to ensure

that the patient has exactly Sickle Anemia, and what is the ratio of effectiveness in the detection process? and what is the important variable of these infected cells that fall in the circle of concentration to give help in the diagnosing decision? These variables are the properties of cells (benign and Sickle), which are computed through the detection process. They are assigned as input variables to applying NN and C&R tree to stand upon the ratio of errors according to the presented algorithm. The input variables consist of Areas, Eccentricities, Perimeters, and Convex areas of all cells, and the target is computed as in Equation (5) depending on the solidity in Equation (4). In the previous, the back propagation has been trained 128 values as an 80% of all samples, 10% validation and testing (i.e. 16 samples for each one).

Experimentally, NN consists of four input variables, ten neurons in one hidden layer and one output layer. Actually, the network already has succeeded after 65 iterations out of 1000 as maximum iterations of the epoch, performance 0.0189, gradient 0.0165, and 6 validation checks as in Figure.6 (a). In (b) the best validation performance is shown as 0.00010338 at epoch 59, when the training in a blue line, validation in green, and the test in the red line. In the same context, the mean square error of training, validation, and testing processes are $2.059e-2$, $1.03383e-2$, and $1.17910e-2$, respectively.

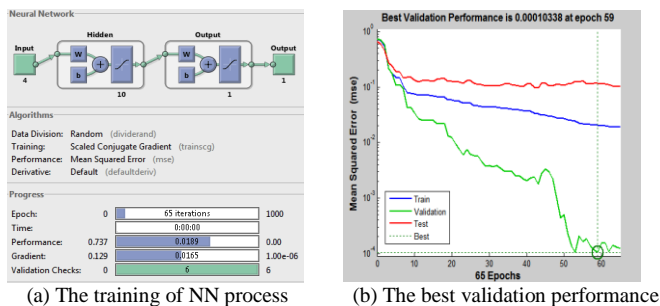


Figure.6 (a) the back propagation NN, and (b) the best validation performance.

Figure.7 shows the confusion matrices for training, validation, and testing processes. In this Figure, the predictions of NN model are compared with the original classes of the target T_c to identify the values of true positives, true negatives, false positives, and false negative. These values have been computed to construct the confusion matrix, where each cell contains the number of cases classified for the corresponding combination of desired and actual classifier outputs, and it achieved 96.9%. Accordingly, accuracy, sensitivity, and specificity are approximated the probability of the positive and negative labels to being true and have assessed the usefulness the algorithm on an NN model. Respectively, the accuracy, sensitivity, and specificity classifications of NN have achieved 98.4%, 100%, and 93.3% success of training samples.

In the same context, the validity has achieved 100%; the accuracy, sensitivity, and specificity classifications of the test samples have achieved 81.3%, 90.9%, and 60.0% respectively.

Last but not least, NN agreed with all detected cells using the proposed algorithm by about 96.9%.

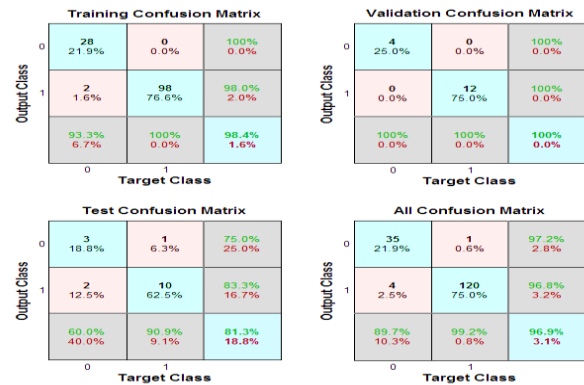


Figure.7 The confusion training, validation, and testing matrices.

The step after is the using of classification and regression tree (C&R), that applies to the same mentioned input and target variables. The tree has carried out with highest depth 5, maximum surrogates 5, and using Gini to measure the impurity. As in the previous, the tree has trained 128 as an 80% of all samples, and 16 samples in validation and testing (i.e. 10% for each one). Actually, the tree has achieved 95.83% in 115 of right predictions (agreed with the target) in the training samples, whereas in the testing samples (cells) achieved 92.86% in 13 of correct agreed with the proposed algorithm target by about 100% in validation. Moreover, the most important and effective determined input variable is the Eccentricity by about 0.6786; the Area variable is coming after by about 0.2576, afterwards Perimeter variable by about 0.0358, and the Convex Area variable at the end by about 0.0279. Therefore, by C&R tree, the diagnosis may depend only on the Eccentricity and then Area variables to distinguish the benign cells from distorted ones.

Furthermore, the accuracy, sensitivity, and specificity in training samples have achieved 89.8%, 100%, and 82.1%, respectively. On the other hand, in the test samples, the accuracy achieved 81.25%, sensitivity has 100%, and the specificity got about 66.7%. Although the C&R tree is easier than NN by applying on those exported data, it has achieved only about 92.9%, while NN has achieved 96.9%. Clearly, the previous applications of NN and C&R tree on the exported data is tending to the decision that NN is preferred and more effective than an C&R tree in prediction on these data. In other words, NN is helped to test and check the efficiency of the data resulted from the proposed algorithm in diagnosing and detection the Sickle and all distorted cells. Finally, Matlab 2013a has been used to build the algorithm on Windows 7 with processor Intel ® Core™2Duo CPU T5550@ 1.83GHz and 2.50 GB RAM with 32-bit Operating system. All the images of blood cell were digitized by the optical Nikon microscope.

8 Conclusions

Microscopic image analysis of human blood cells helps as a diagnostic tool for the infected blood cells' detection.

Sickle-cell anemia is one of the most important common types of anemia disease. This paper has been presented a proposed algorithm that can be detecting and counting the Sickle and all distorted cells in a microscopic colored image; even if they are hidden or overlapped. The algorithm has been used circular Hough transform to detect the benign and distorted blood cells. On the other hand, the exported variables data (Areas, Convex Areas, Eccentricity, and Perimeter) for all detected cells (benign, and distorted) have been classified as input variables, whereas a solidity measure variable for all of cells has been constructed as a target variable. In the next step, the neural network and regression tree have been applied to get the right decision for diagnoses and check the effectiveness of the proposed algorithm in detection. The experimental results have been demonstrated high accuracies and success of these models in predicting the infected cells that are contained Sickle Anemia or distorted cells. The performances have been calculated by three statistical measures; classification accuracy, sensitivity, and specificity. This algorithm has been achieved in segmentation and classification processes by about 99.98% out of all input cells are detected, which may have contributed to improve diagnosis of Sickle Anemia diseases. The experimental results have been shown that the effectiveness reaches to 96.9% in the case of applying Neural network and 92.9% when using C&R tree. Therefore, the proposed algorithm is very effective in the detection of benign and distorted red blood cells, additionally that the neural network is more efficient than a C&R tree in the case of testing the quality of the algorithm in detection.

9 References

- [1]. B. Chaudhuri and U. Bhattacharya." Efficient training and improved performance of multilayer perceptron in pattern classification." *Neuro computing*, 34, pp11–27, September 2000.
- [2]. B. R. Devi, K. N. Rao, S. P. Setty and M. N. Rao. Disaster Prediction System Using IBM SPSS Data Mining Tool. *International Journal of Engineering Trends and Technology (IJETT)*-Volume4 Issue8-August 2013 ISSN Volume: 2231.
- [3]. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. *Classification and Regression Trees*, Belmont, California: Wadsworth, Inc, 1998.
- [4]. Freeman WJ. "The physiology of perception." *University of California, Berkeley. SCI Am.* 1991 Feb;264(2):78-85.
- [5]. H. A. Khan and G. M. Maruf. Counting clustered cells using distance mapping. *International Conference on Informatics, Electronics & Vision (ICIEV)*, P:1-6, 17-18 May 2013.
- [6]. H. Fleyeh, R. Biswas and E. Davami. Traffic sign detection based on AdaBoost color segmentation and SVM classification EUROCON conference, IEEE. 2005-2010. 1-4 July 2013.
- [7]. Hany. A. Elsalamony, Alaa. M. Elsayad. Bank Direct Marketing Based on Neural Network. *International Journal of Engineering and Advanced Technology IJEAT*, ISSN: 2249–8958, Vol-2, Issue-6, Aug 2013.
- [8]. Image Processing Toolbox, is available through MATLAB's help menu, or online at: <http://www.mathworks.com/help/images/index.html>
- [9]. K. Thirusittampalam, M. J. Hossain, O. Ghita, and P. F. Whelan. A Novel Framework for Cellular Tracking and Mitosis Detection in Dense Phase Contrast Microscopy Images. *IEEE Journal of Biomedical and Health Informatics*, Vol. 17, No. 3, MAY 2013.
- [10]. M. C. Mushabe, R. Dendere and T. S. Douglas. Automated detection of malaria in Giemsa-stained thin blood smears. *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC*, P: 3698-3701, 3-7 July 2013.
- [11]. S. D. Brown and A. J. Myles. *Decision Tree Modeling in Classification*. Book chapter Editor: S. D. Brown, R. Tauler and B. Walczak, Book Title: *Comprehensive Chemometrics*. Elsevier. Pages: 541-569. 2009.
- [12]. Stockman GC, Agrawala AK. Equivalence of Hough curve detection to template matching. *Communications of the ACM.* 1977;20:820-2. 20:820-2.
- [13]. T. J. Atherton and D. J. Kerbyson. Size invariant circle detection. *Image and Vision computing*. Volume: 17. Issue: 11 1999. Pages: 795-803.
- [14]. TIAN YuBo, ZHANG XiaoQiu, and ZHU RenJie. "Design of Waveguide Matched Load Based on Multilayer Perceptron Neural Network." *Proceedings of ISAP, Niigata, Japan 2007*.
- [15]. Tom M. Mitchell. "Machine Learning." *Learning.* Copyright c 2005-2010. Second edition of the textbook. Ch 1. January 19, 2010 McGraw Hill. 119 ACM 2005, ISBN 1-59593-180-5.
- [16]. U. S. Department of Health & Human Services, National Institute of Health. <http://www.nhlbi.nih.gov/health/health-pics/topics/sca/>
- [17]. Wikipedia, the free encyclopedia, (Redirected from Neural network). http://en.wikipedia.org/wiki/Neural_network#History_of_the_neural_network_analogy

Estimation of Resected Liver Regions Using a Tumor Domination Ratio

Masanori Hariyama¹, Moe Okada¹, Mitsugi Shimoda², Keiichi Kubota²

¹Graduate School of Information Sciences, Tohoku University, Japan

² Second Department of Surgery, Dokkyo Medical University

Abstract—This article presents an automatic approach to estimate optimal resected-liver regions for oncologic surgery planning. Usually, resected liver regions are determined by selecting cut points on the portal vessels on 3D simulation software. Since the liver has complex vessel structure, it is difficult for human to find optimal resected liver regions. To solve this problem, a tumor domination ratio is proposed to find all portal vessels related to tumors. The tumor domination ratio allows us to compute the ideal resected region, that is, all the perfusion territories related to the tumor. Moreover, some types of conditions such as the size of vessels are considered for practical surgical use. The experimental results demonstrate that the resected liver regions of the proposed approach are much smaller than those of the conventional approach in most cases.

Keywords: Medical imaging, 3D simulation analysis, anatomic hepatectomys

1. Introduction

3D simulation, recently, plays an important role in surgical planning for hepatectomy since the liver has a complex structure, that is, some different vessels are arranged complexly as shown in Fig. 1. The estimation of regions perfused by the portal vein is especially one important task in anatomical hepatectomy, since the hepatocellular carcinoma (HCC) tends to metastasize via the portal vein [1], [2]. Figure 2 explains how the perfused region is estimated. The tumor affects the nearest parts of the portal vein; from the nearest parts, the HCC metastasize in the downstream direction via the portal vein. The overestimate of the perfused region tends to prevent the metastases, but it increases the possibility of the postoperative liver failure. Therefore, the volume of perfused region should be minimized while including all the perfused regions of the portal vein feeding the tumor.[2]

Some practical software programs for 3D simulation and preoperative planning have been developed like Hepavision (MeVis Medical Solutions AG, Bremen, Germany)[3], OVA (Hitachi, Tokyo, Japan)[4], and Synapse VINCENT (Fuji-film Medical, Tokyo, Japan)[5]. In preoperative planning, surgeons search for the combination of the cut points on the portal vein in a trial-and-error way until the tumor is covered by the perfused regions of the portal cut points. However, it is difficult and time-consuming for surgeons to

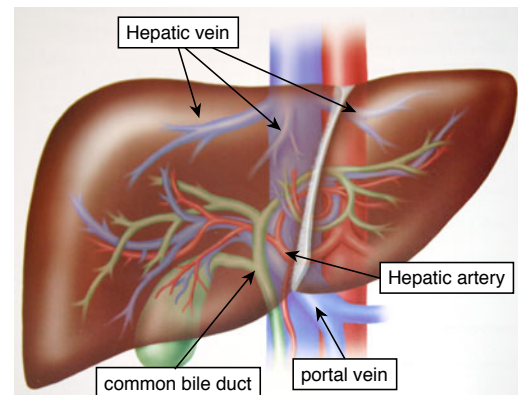


Fig. 1: Vascular system of the liver.

find optimal or near-optimal combination since the portal vein has complex structure, that is, it has many branches. Surgeons select simple combinations, and this results in overestimation of the perfused region.

To solve this problem, this paper presents the method to compute the optimal perfused regions, where the volume is minimized while including all the perfused region of the tumor-related portal vein. In order to find the tumor-related part of the portal vein accurately, the tumor domination ratio is defined which reflects how much volume of the tumor a part of the portal vein feeds. The tumor domination ratio allows to pick up all the tumor-related parts of the portal vein, and the total of the regions perfused by them is determined to be the resected region.

The rest of this paper is organized as follows. In Section 2, the basic method to find the ideal resected region is explained where practical conditions used in the real surgeries are not considered. In Section 3, its extension is explained where practical conditions such as the branch points and the radiuses of the vessels are considered. Section 4 is conclusion.

2. Estimating an ideal resected region using the tumor donation ratio

The region perfused by a point on the portal vein is estimated by using a Voronoi diagram[2]. Figure 3 shows a Voronoi diagram. A Voronoi diagram is a way of dividing

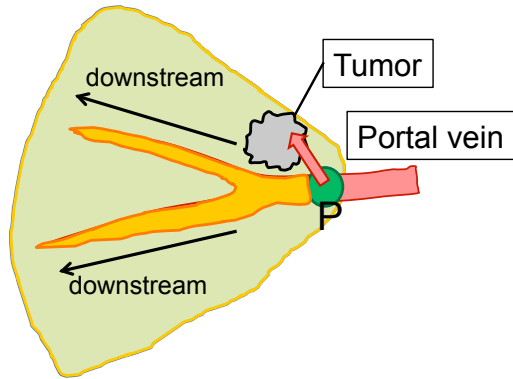


Fig. 2: Region perfused by the portal point P and its downstream.

space into a number of regions. A set of points (called seeds) is specified beforehand and for each seed there will be a corresponding region consisting of all points closer to that seed than to any other. The regions are called Voronoi cells. A seed and Voronoi cell in a Voronoi diagram correspond to a point on a portal vein and a region perfused by the portal point.

The tumor domination ratio TDR of a portal point P is defined as follows.

$$TDR = \frac{[\text{Volume of the region perfused by } P]}{[\text{Volume of the tumor}]} \times 100 \quad [\%] \quad (1)$$

Given a tumor location, let us compute an ideal resected region by using the tumor domination ratio, where the ideal resected region means a minimum region including all sub-regions perfused by tumor-related portal points. All tumor-related portal points are found by using the tumor domination ratio; all the portal point with TDR larger than 0 are picked up as tumor-related portal points as shown in Fig. 4. The ideal resected region is given as the union of all the regions perfused by the tumor-related portal points.

Let us compare the proposed method with the manual approach by the conventional 3D simulation software program (Synapse VINCENT ver.2). Figure 5 shows the sample data used for evaluation. The samples 1 and 2 are the cases where the small tumor exists near the liver surface; the sample 3 is the case where the large tumor exists near the main stem of the portal vein. For each sample, the liver, vessels, and a tumor are extracted in advance by the other software programs. Tables 1-3 summarize the results for sample 1-3, respectively. From Tables 1 and 2, the proposed optimization method is very useful compared to the manual approach, that is, the volume of the estimated resected region by the proposed method is much smaller than that of the manual approach. This is because the degree of freedom of cut-point selection is high and the effect of optimization is large when the case where the small tumor exists near the liver surface. On the other hand, from Table 3, the

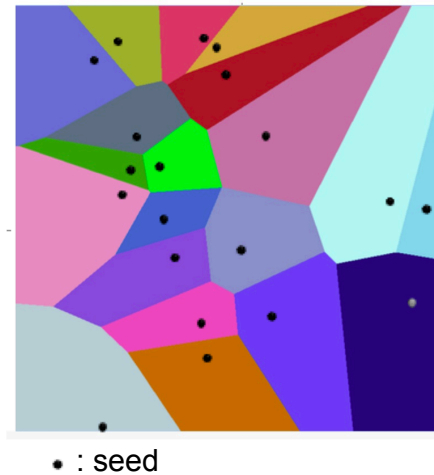


Fig. 3: Voronoi diagram.

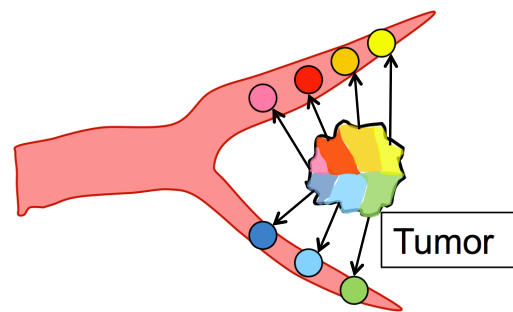


Fig. 4: Tumor-related portal points.

optimization is not very effective when the tumor exists near the main stem of the portal vein since the degree of freedom of cut-point selection is low.

3. Estimating an optimal resected region under the practical conditions in surgery

Although the method described in the previous section greatly reduces the volume of the estimated resected region, it is difficult for surgeons to use the results directly in the real surgery since the number of cut points is larger than that of manual approach, and since cut points are sometimes set on such tiny vessels that surgeons cannot identify in the real surgery. In order to solve this problem, the the method described in the previous section is improved considering the practical conditions as follows. In this paper, branch points and the radius of the vessels are considered as the practical conditions. The optimization procedure starts with obtaining the ideal cut by the method in the previous section shown in Figure 6(a). These ideal cut points are refined considering the practical conditions as shown in 6(b). Each cut point is moved towards the upstream direction in such a way that the cut point is set on a branch and the radius of the vessel

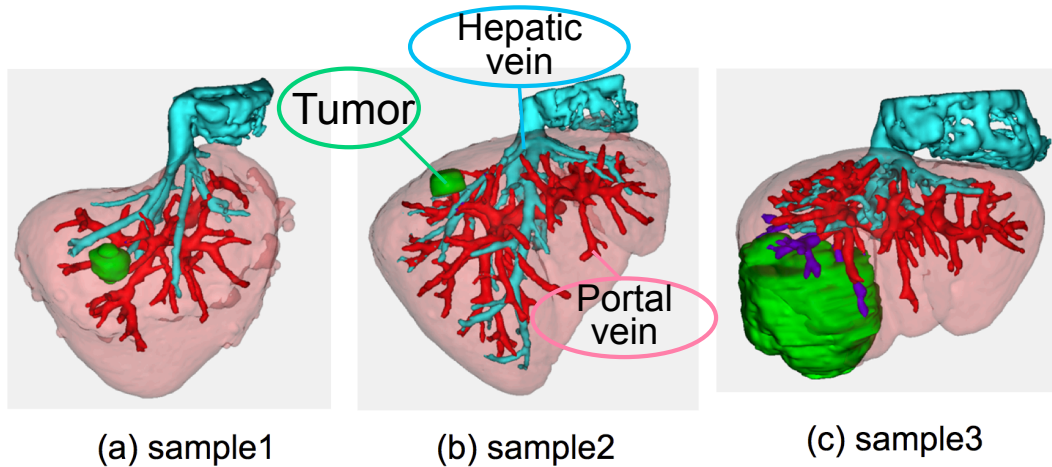


Fig. 5: Samples for evaluation.

Table 1: Comparison results for sample 1

sample1	Manual approach by a surgeon	Proposed
Resected volume (cc)	192	66
The number of cut points	1	2
Results		

Table 2: Comparison results for sample 2

sample2	Manual approach by a surgeon	Proposed
Resected volume (cc)	325	42
The number of cut points	2	4
Results		

Table 3: Comparison results for sample 3

sample3	Manual approach by a surgeon	Proposed
Resected volume (cc)	550	476
The number of cut points	5	10
Results		

Table 4: Results of the optimization considering the practical conditions for sample 1.

sample1	Manual approach by a surgeon	Proposed (Radius: 6mm)	Proposed (Radius: 8mm)
Resected volume (cc)	192	112	178
The number of cut points	1	2	1
Results			

Table 5: Results of the optimization considering the practical conditions for sample 2.

sample2	Manual approach by a surgeon	Proposed (Radius 6mm)	Proposed (Radius 8mm)
Resected volume (cc)	325	170	346
The number of cut points	2	3	2
Results			

is larger than the pre-determined condition.

Figures 4 and 5 summarize the results of the optimization

considering the practical conditions for samples 1 and 2, respectively. The radius of the vessel is set to 6 mm and

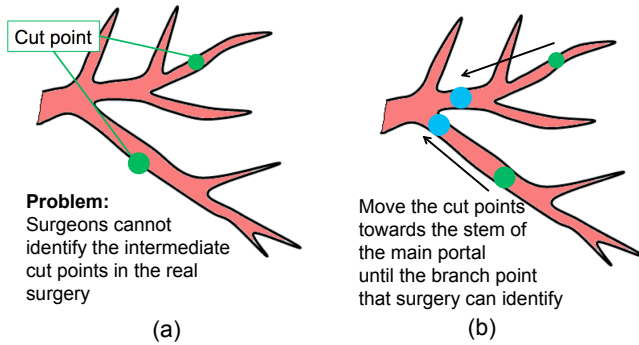


Fig. 6: Refining the cut points considering the practical conditions.

8 mm. These results shows that the estimated resected region is minimized keeping the number of cut points small appropriately for real surgery

4. Conclusion

In order to minimize the resected volume while preventing metastases, the automatic method that finds the optimal combination of cut points is proposed.

The domination ratio is key to success to find such optimal combination in a systematic way.

Recently, the integration of 3D simulation and ultrasonography is progressing to support the surgical operations in the real surgery. Such navigation technology makes it easier that surgeons resects the liver according to the optimal resected regions computed from our method.

References

- [1] M. Makuuchi, H. Hasegawa, S. Yamazaki, "Ultrasonically guided subsegmentectomy", *Surg. Gynecol. Obstet.*, Vol.161,pp.346-50 (1985).
- [2] T. Takamoto, T. Hashimoto, S. Ogata, K. Inoue, Y. Maruyama, A. Miyazaki, M. Makuuchi, "Planning of anatomical liver segmentectomy and subsegmentectomy with 3-dimensional simulation software", *The American Journal of Surgery* Vol. 206, Issue 4, pp. 530-538 (2003)
- [3] D. Selle, B. Preim, A. Schenk, et al., "Analysis of vasculature for liver surgical planning", *IEEE Trans. Med. Imaging*, Vol. 21, pp.1344-57 (2002).
- [4] S. Saito, J. Yamanaka J,K. Miura, et al., "A novel 3D hepatectomy simulation based on liver circulation: application to liver resection and transplantation", *Hepatology*, Vol. 41, pp.1297-1304, (2005)
- [5] S. Ohshima, "Volume analyzer SYNAPSE VINCENT for liver analysis", *Journal of Hepato-Biliary-Pancreatic Sciences*, Vol. 21, Issue 4, pp. 235-238(2014)

Rebuilding IVUS Images From Raw Data Of The RF Signal Exported by IVUS Equipment

Marco Aurélio Granero^{1,3}, Marco Antônio Gutierrez², Eduardo Tavares Costa¹

¹ Department of Biomedical Engineering– DEB/FEEC/UNICAMP, Campinas, Brazil

² Division of Informatics/Heart Institute – HCFMUSP, São Paulo, Brazil

³ Federal Institute of Education, Science and Technology S. Paulo – IFSP, São Paulo, Bra

Abstract - *The study of composition and classification of atherosclerotic plaque has been a very active research field, both in cardiology and image processing. Intravascular ultrasound (IVUS) is an effective tool, which can insights about the cross-section of blood vessels, with sufficient accuracy to allow an accurate assessment of CT slices. This enables information about blood vessel structures to be determined. During an IVUS medical examination, physicians subjectively adjust a set of parameters to improve the visualization of a Region Of Interest (ROI) and produce corresponding images in Digital Imaging and Communications (DICOM) format, for later analysis and study. DICOM is appropriate for storage, transportation and access, but limits subsequent changes to image parameters, such as contrast or brightness. This makes comparison across patient populations difficult and restricts image processing operations. This paper details an alternative to using DICOM, which is to rebuild IVUS images from raw radiofrequency signal (RF) data. The main advantage of this process is the independence of the acquisition parameters adjusted during the exam. This advantage makes possible the comparison between exams and can be used to monitor the evolution of cardiovascular disease. Beyond this, once the reconstructed images and the RF signal are stored, operations relating to texture and spectral analysis can be carried out and automatic classifiers employed. From a clinical point of view these reconstructed images share the same characteristics as DICOM images with an advantage that the former have a higher contrast than the latter, allowing deeper regions to be seen.*

Keywords: RF signal, IVUS image, rebuilding process, ultrasound image, RF raw data.

1. Introduction

Among the different modalities of medical images, ultrasound is arguably the most difficult in which to perform segmentation. This is evident from a study of the first papers on segmentation, in which it was only possible to apply a threshold to the image in order to separate the background from foreground due to the poor quality of the acquired data (Noble, 2010).

At the same time, subsequent technological development has greatly increased the quality of ultrasound images, especially in terms of signal to noise ratio (SNR) and contrast to noise ratio (CNR), resulting in improvements to image quality. Several studies have been highlighted that aim to develop algorithms for the design of edges on objects contained in ultrasound images (Noble, 2010).

Ultrasonic Tissue Characterization (UTC) has become a well-established research field since its first publication (Mountford and Wells, 1972). Thijssen (2003) defines UTC as the assessment by ultrasound of quantitative information about the characteristics of biological tissue, and their pathology. This quantitative information is extracted from echographic data from RF data.

UTC applications abound in the literature and include classification of breast tissue (Tsui et al. 2008 and Molthen et. al. 1998), liver (Molthen et al., 1998), heart (Clifford et al., 1998 and Nillesen et. al., 2008), eyes (Lizzi et. al., 1983), skin (Raju et. al., 2003), kidney (Engelhorn et. al., 2012) and prostate (Moradi, 2008).

Szabo (2004) defines two general goals for ultrasonic tissue characterization which can be applied to the above areas (Szabo, 2004):

- i. Reveal the properties of tissues by analyzing the RF signal backscattered by ultrasound transducer and
- ii. Use information about the properties of the tissue to distinguish between the state of tissue (healthy or diseased), or to detect changes in these properties when subjected to stimuli or long periods of time in response to natural processes or medication.

Reaching these goals can be challenging since the interaction between biological tissue and sound waves is extremely difficult to model and the process evolved in image segmentation is strongly influenced by the quality of data and by the different parameters used during the acquisition process of an image.

Parameters like contrast, brightness and gain are adjusted by physicians to improve the visualization of regions during the examination. These changes determine the DICOM images that are recorded and the result cannot be changed after the image has been acquired. This greatly complicates the comparison between patients and the use of images in studies of groups of patients.

Thus, to avoid these complications and make image reconstructed independent of the parameters set by the physician a reconstruction method from IVUS images is proposed. This method is based on the RF signal stored by the equipment during medical imaging examinations of intravascular ultrasound.

The process of rebuilding starts with applying a band-pass filter to the RF signal to eliminate signals that do not come from the transducer. In the next step, a time gain compensation (TGC) function is applied to compensate for attenuation loss. After this, the envelope of the signal is computed and the result is log-compressed and normalized in a grayscale image.

After the process of rebuilding, the grayscale image, in polar coordinators, is submitted to a Digital Development Process (DDP) responsible for enhancing the contrast and edge emphasis. So, the image is interpolated to cartesian coordinators. The cartesian image is further processed with an intensity transformation function to improve the contrast of the final cartesian grayscale IVUS image.

The above processes are described in more detail in section 2 and the results obtained are shown in section 3. In section 4 a comparison is made between reconstructed images and DICOM images from an examination. Finally, section 5 shows conclusions and possibilities for future work.

2. Method for IVUS image reconstruction

An IVUS examination is carried out by inserting a catheter into coronary arteries via femoral or brachial vessels. At the tip of this catheter there is an ultrasound emitter and a piezoelectric transducer that collect the echoes reflected by internal structures of the vessel as RF signal.

A schematic representation of the execution of an IVUS examination is shown in Figure 1(a), where the IVUS equipment collects data from patient and stores it in the workstation. Figure 1(b) shows an IVUS rotational catheter.

During an IVUS exam, the acquired images are stored in DICOM format and exported to the databank of the clinical centre to be used for clinical diagnosis.

In addition to the images in DICOM format, the equipment allows the RF signal to be recorded, which are used in the manufacture of images in a proprietary format.

The proposal of this paper is to process the RF signal data according to the steps shown in Figure 2. These steps are detailed below.

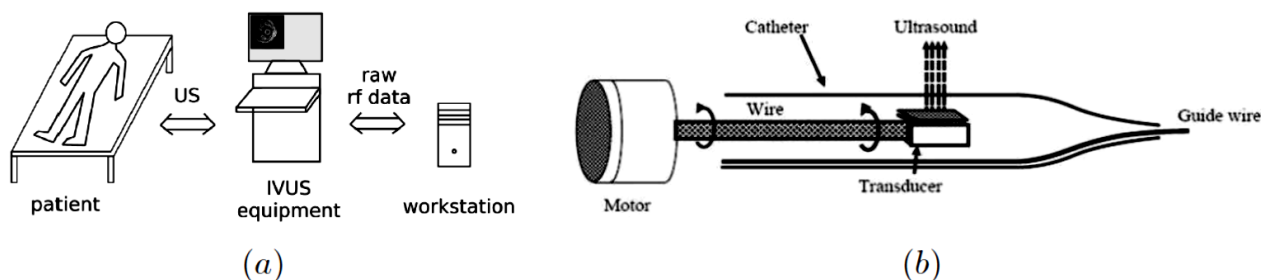


Figure 1: (a) IVUS *in-vivo* analysis typical scenario, (b) Rotational IVUS catheter. Extracted from Ciompi, (2008).

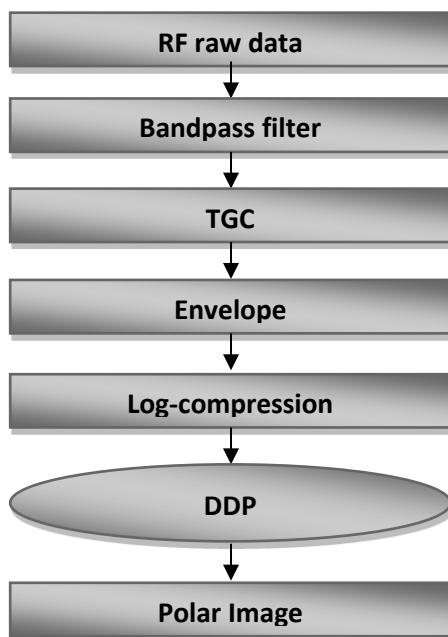


Figure 2: Block diagram of reconstruction process.

2.1 – RF dataset

The data was taken from examinations in the Department of Hemodynamics in the Heart Institute of the Medical School of the University of São Paulo (Heart Institute – HCFMUSP), Brazil, using iLab IVUS (Boston Scientific, Fremont, USA), equipped with a 40 MHz catheter Atlantis SR 40 Pro and anonymised to avoid the identification of the patient and used only for research purposes.

The RF File Reader (designed by Boston Equipment) is an xml file that contains information about the examination. This file allows us to identify the number of rows, columns and frames from each exam. Beyond this, the reader contains the distance from each pixel in the image, in millimetres.

Once image attributes have been found using the RF File Reader, it is possible to extract the data. These data were placed in a tri-dimensional matrix. The rows of this matrix represent the lines in A-mode, each line with radial information about the vessel, the columns represent the distance to the tip of catheter and the slices, third dimension, represent each time frames of the exam. The study of IVUS used in this work results

in a 3D matrix, where the dimensions represents the size of each image and the third dimension being the number of frames.

After this, each frame was submitted to the reconstruction shown in Figure 2.

2.2 - Bandpass filter

A Butterworth bandpass filter was applied to dataset in order to eliminate frequencies that do not come from the transducer. The manufacture of transducer describes the central frequency emitted by transducer at the tip of catheter as 40 MHz and frequency sample rate as 200 MHz.

Each line in A-mode was filtered by a Butterworth finite impulse response filter (FIR filter).

The frequency range was adjusted between 20 and 60 MHz as can be viewed in Figure 3(a).

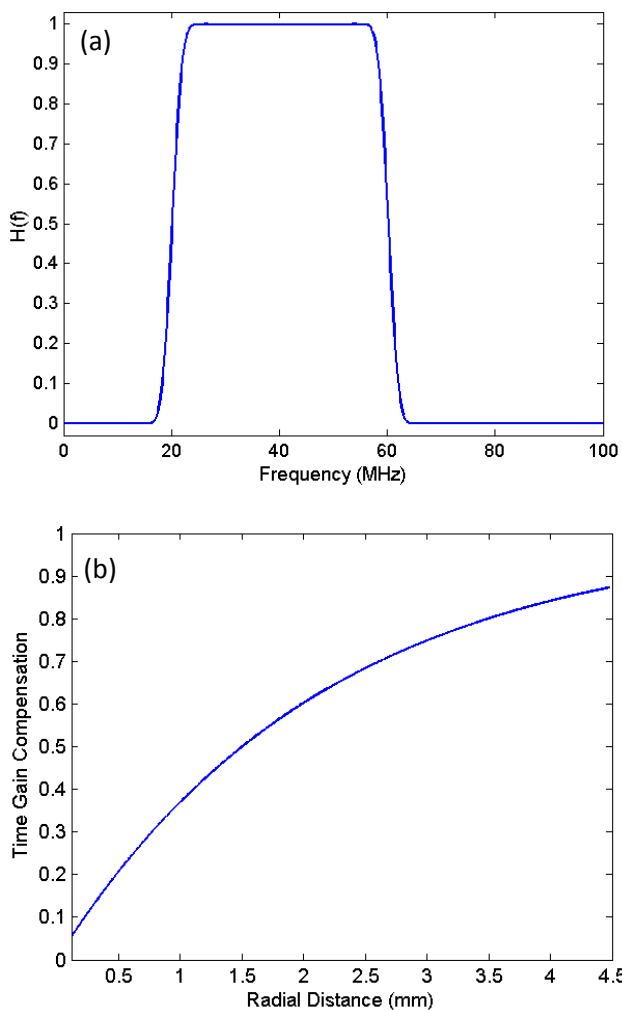


Figure 3: (a) Frequency response of the Butterworth FIR filter and (b) Profile of TGC function.

2.3 - TGC (Time Gain Compensation)

The ultrasound beam is attenuated as it penetrates the tissue. To compensate for this loss in signal intensity TGC is applied to each line in A-mode, which is defined as

$$T(r) = 1 - e^{-\beta r} \quad (1)$$

where β is the coefficient of attenuation and r is the radial distance from tip of catheter.

The range of the radial distance was extracted from the RF File Reader of exam ranging until 4.48cm.

In Ciompi (2008), RF signal of in-vivo and ex-vivo was used to develop a multiclass classifier to the problem of characterization of the atherosclerotic plaque. They define a value for the coefficient of attenuation as $\beta = 0,4605$ dB/cm, which was adopted in this work.

The profile of TGC function is shown in Figure 3(b).

2.4 - Signal envelope

To show the changes stemming from the texture and not from the wave profile, the envelope of the signal is obtained simply applying the Hilbert transform to each line in A-mode from the RF signal.

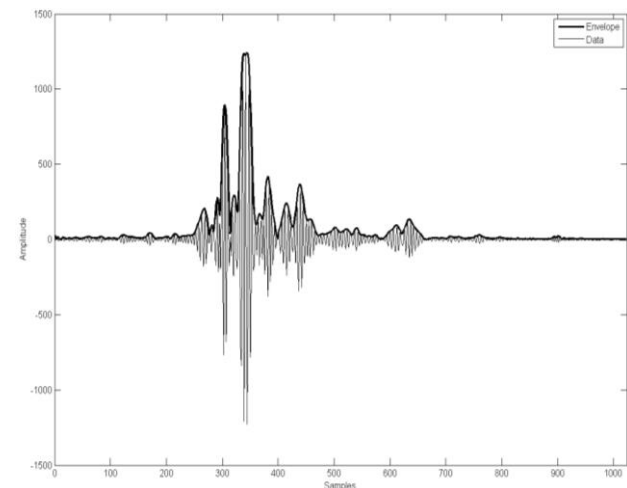


Figure 4: RF envelope is shown as a black line over the wave profile of RF signal gray line (Data).

2.5 - Log-compression

The next stage in the procedure described in Figure 2, is to normalize the RF signal providing values between 0 and 1 in order to permit work with a homogeneous range for all IVUS images. After this, the RF signal undergoes a transformation whose purpose is to map a narrow range of grayscale values in an input image to a wide range of output levels (Gonzalez and Woods, 2004). This transformation is defined as

$$I_{\log} = \frac{1}{t} \log[1 + (e^t - 1)I_{nor}] \quad (2)$$

where I_{nor} represents the RF signal normalized and t is empirically obtained to improve the log-compression.

2.6 - Digital Development Process

In order to emphasize the edges borders and improve contrast gain, a Digital Development Process (DDP) (Gonzalez and Woods, 2004) was applied to the RF signal.

Each pixel value of an image was modified by equation (3) to produce an image with better Contrast Noise Ratio (CRN).

$$Y_{ij} = k \left(\frac{X_{ij}}{\{X_{ij}\} + a} \right) + b \quad (3)$$

where $\{X_{ij}\}$ is obtained by applying a Gaussian low-pass filter to the original image and the parameters k , a , and b were empirically determined to improve the CRN.

After this, the image was converted and interpolated to cartesian form, resulting in an image with 512x512 pixels and 256 gray levels.

Finally, an intensity transformation was applied to image in order to expand the saturation of the gray level dynamic band and the image was Gaussian filtered.

3. Experimental Results

The results of the rebuilding process of the IVUS images are shown in Figure 5, which the mayor structures visible in an IVUS examination are identified.

Figure 5(a) shows the segmentation of lumen and the media-adventitia borders. 5(b) the stent and an artifact generated by the wire guide. 5(c) shows a region with calcification and the acoustic shadow behind it, with an arrow pointing to an artifact generated by the wire guide.

Figures 5(e) and (h) show a bifurcation region, with calcification. A stent is visible in 5(d) and (i), and it is possible to identify the malposition of the stent in 5(i).

Figure 5(f) shows the shadow of the pericardium and 5(g) the acoustic shadow of a big calcification and the lumen and media-adventitia borders.

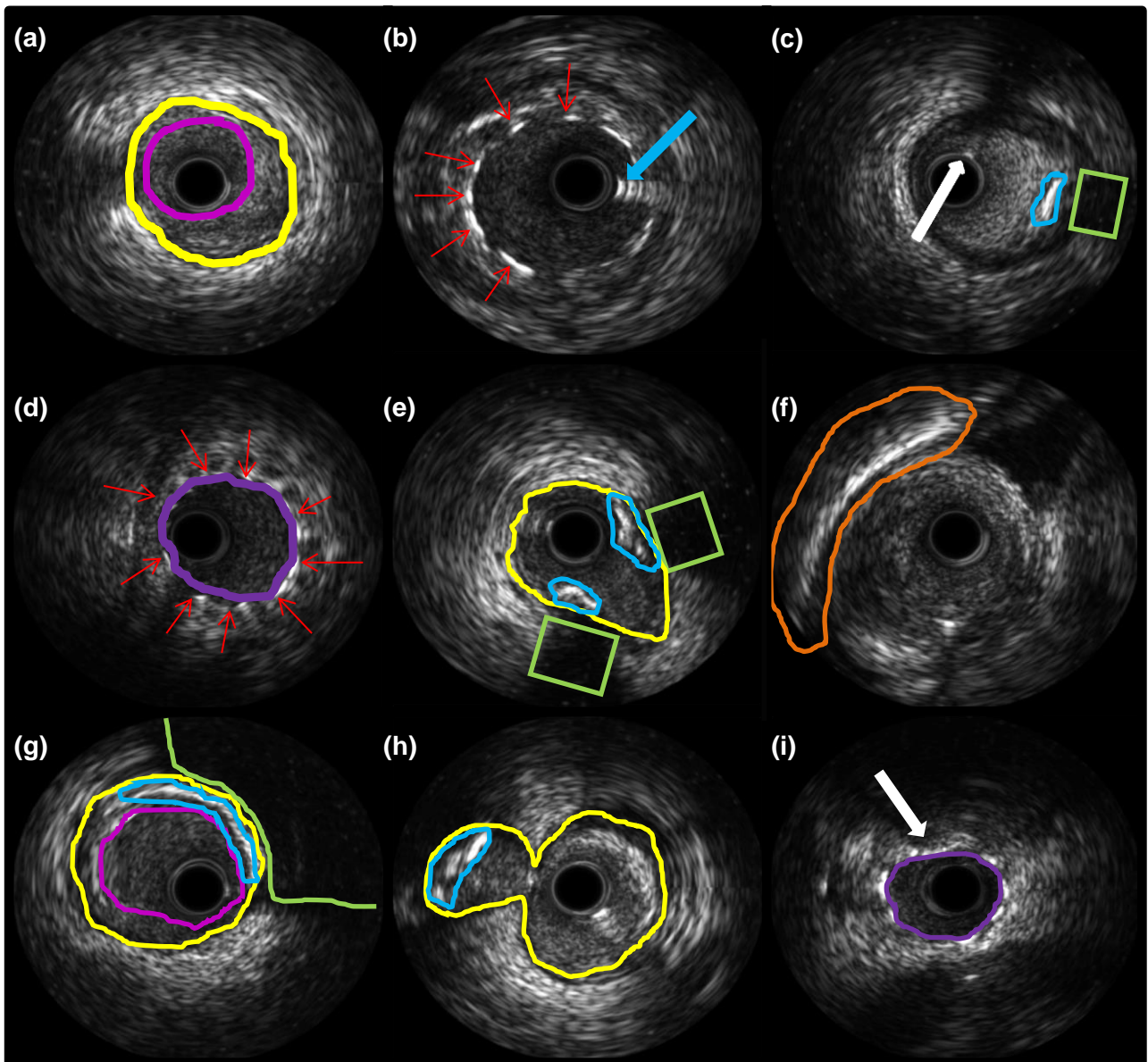


Figure 5: Images from the reconstruction process.

Figure 6 show both the rebuild image and the DICOM images.

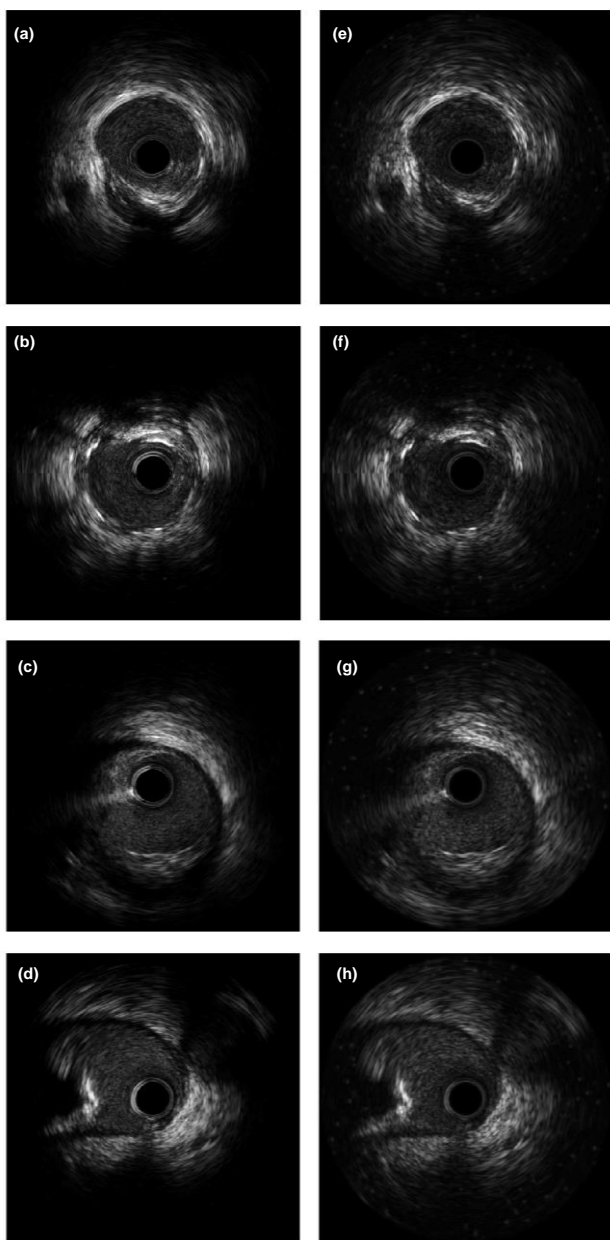


Figure 6: (a), (b), (c) and (d) DICOM images.
(e), (f), (g) and (h) Rebuilt images.

As can be seen, the rebuilt images show the same structures as DICOM images, and in all cases the contrast of the rebuilt images are better than DICOM images. What is perhaps most noticeable is the difference in visibility in the outer region of the lumen. The reconstructed image shows fine detail where the DICOM shows only a black region.

4. Discussion, Conclusion and Future Work

IVUS is an examination that can provide a good quality image of the cross-section of blood vessel allowing the assessment of inner structures.

In an IVUS medical examination, sets of hundreds or even thousands of images are acquired and used as the basis for a medical diagnosis.

These images are subject to a variability of interpretation inter and intra operator because a set of parameter are adjusted to improve the visualization of a ROI. Once the images are acquired these parameters cannot be changed, restricting the comparison between different examinations or patients.

To avoid this limitation, this article describes a methodology for reconstructing IVUS images from RF raw data, which are independent of the parameters adjusted by the physician during the exam and which can be processed to improve the CNR of the image.

The RF signal is processed according to the theoretic model proposed in section 2 and illustrated in Figure 2. The parameters used in the model were adjusted to maximize CNR enabling identification of the main structures of the vessel.

The results of the proposed model were presented in Figures 5 and 6 and compared with the DICOM images generated by the equipment. The proposed model produces images with superior CNR which can be used for clinical purposes.

In the figures it is possible to see the main structures of the vessel and this result can be used to perform segmentation to help the physician in diagnosis process. Beyond this, it is possible to identify bifurcations and calcifications regions to be submitted a percutaneous coronary intervention - PCI.

Considering the data used in this work, the propose method was proved to be robust with regard to fidelity in the reconstruction of structures in comparison with DICOM image and, in all cases the CNR in reconstructed images was greater than DICOM images, figure 6.

5. References

- [1] Noble, J. A., "Ultrasound image segmentation and tissue characterization", Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine vol. 224: 307, 2010. DOI: 10.1243/09544119JEIM604
- [2] Mountford, R. A. and Wells, P. N. T. (1972). "Ultrasonic liver scanning: The A-scan in the normal and cirrhosis". Phys. in Med. & Biol. 17, 261–269.
- [3] Thijssen, J. M., "Ultrasonic speckle formation, analysis and processing applied to tissue characterization", Pattern Recognition Letters, vol. 24, 659-675, 2003

- [4] Tsui, P.-H., Yeh, C.-K., Chang, C.-C., and Liao, Y.-Y. "Classification of breast masses by ultrasonic Nakagami imaging: a feasibility study", *Phys. Med. Biol.*, 53, 6027-6044, 2008.
- [5] Molthen, R. C., Shankar, P.M., Reid, J.M., Forsberg, F., Halpern, E. J., Piccoli, C. W., and Goldberg, B. B. "Comparisons of the Rayleigh and K distribution models using in vivo breast and liver tissue", *Ultrasound in Medicine and Biology*, 24, 93–100, 1998.
- [6] Clifford, L., Fitzgerald, P., and James, D. "Non-Rayleigh first-order statistics of ultrasonic backscatter from normal myocardium", *Ultrasound in Medicine and Biology*, 19, 487-495, 1993.
- [7] Nillesen, M. M., Lopata, R. G. P., Gerrits, I. H., Kapusta, L., Thijssen, J. M., and de Korte, C. L., "Modeling envelope statistics of blood and myocardium for segmentation of echocardiographic images". *Ultrasound in Medicine and Biology*, 34(4), 674–680, 2008.
- [8] Lizzi, F. L., Greenbaum, M., Feleppa, E. J., Elbaum, M., and Coleman, D. J., "Theoretical framework for spectrum analysis in ultrasonic tissue characterization", *Journal of Acoustic Society America*, 73, 1366–1373, 1983.
- [9] Raju, B. I., Swindells, K. J., Gonzalez, S., and Srinivasan, M. A., "Quantitative ultrasonic methods for characterization of skin lesions in vivo", *Ultrasound in Medicine and Biology*, 29(6), 825–838, 2003.
- [10] Engelhorn, A. L. D. V., Engelhorn, C. A., Salles-Cunha, S. X., Ehlert, R., Akiyoshi, F. K. e Assad, K. W., "Ultrasound tissue characterization of the normal kidney", *Ultrasound Quarterly*, vol. 28, n° 4, December 2012.
- [11] Moradi, M., "A New paradigm for Ultrasound-Based Tissue Typing in Prostate Cancer". Tese de doutorado. School of Computing, Queen's University. 2008.
- [12] Szabo, T. L., *Diagnostic Ultrasound Imaging Inside Out*, Hartford, Connecticut, Elsevier, 2004.
- [13] Ciompi, F. "Ecoc-based plaque classification using in-vivo and ex-vivo intravascular ultrasound data". Master thesis. Computer Vision Center. Universitat Autònoma de Barcelona. 2008.
- [14] Gonzalez, R. C., Woods, R. E. e Eddins, S. L. *Digital Image Processing Using Matlab*. Prentice Hall. 2004.

ACKNOWLEDGEMENTS

This work is supported by the Brazilian National Institute of Science and Technology in Medicine Assisted by Scientific Computing (INCT - MAAC) and National Council for Scientific and Technological Development (CNPq).

Image Segmentation Techniques Applied to Point Clouds of Dental Models with an Improvement in Semi-Automatic Teeth Segmentation

Tamayo-Quintero, J. D.¹, Arboleda-Duque, S.^{1,2} and Gómez-Mendoza, J.B.¹

¹Department of Electric, Electronic and Computer Engineering, Universidad Nacional de Colombia, Manizales, Caldas, Colombia

²Department of Telecommunication Engineering, Universidad Catolica de Manizales, Manizales, Caldas, Colombia

Abstract - This paper presents an exploratory study on the application of a combination of different segmentation techniques to point clouds of dental models. The techniques are based in geometric primitives (e.g. RANSAC), region growing segmentation and graph theory (particularly the "Min-Cut" algorithm), and were tested using dental 3D point clouds. Data were acquired using a Konica Minolta Vivid 9i laser range scanner.

Also, a semi-automatic segmentation methodology is presented. Results of teeth segmentation using testing data suggest that it is possible to automatically segment teeth from digital 3D models.

Keywords: Point Cloud, 3D dental models, segmentation, region growing, RANSAC, Min-Cut.

1 Introduction

The advent of new technologies has driven the development of increasingly sophisticated 3D acquisition systems [1]. 3D models originated by using those systems, also referred as point clouds, provide surface information, metrics, texture, etc. This 3D information is of great interest in different fields in the industry, in areas like in-process inspection, accident reconstruction, crime scene analysis, machine calibration, orthodontics, etc. [2].

Digital dental models have proved to be helpful and important for experts in odontology and orthodontics. They provide information precise enough to be used in diagnosis and prognosis [3], thus using 3D dental models is becoming a more common practice in the field.

Colombian laws in consumer protection requires dental cast records of patients to be preserved during at least 10 years, and according to the Association of Orthodontists [4], it is recommended that study models are retained for at least 11 years or until the patient is 26 years old. This involves a number of concerns to emerge, regarding limited storage capacity, model fragility, high economic costs, etc.

Because of those concerns, it is important to use the 3D digital dental models in place of physical alginate casts. Aiming to further strengthen the acceptance of such practice, 3D image processing and digital measuring tools are finding their way into orthodontics.

Surface and object segmentation is an intermediate step in artificial vision that eases object recognition and classification. Therefore, a good segmentation is vital to facilitate, enhance, and achieve further interpretation of the input data.

For that reason, in this work we present an exploratory analysis of current segmentation techniques for point clouds applied to dental 3D models. This is a first step towards parameter measurement [5] [6], simulation of the movement of teeth to correct malocclusions [7], planning for dental and maxillofacial surgery [8], pose estimation [9], among others.

This paper is organized as follows: Section 2 describes briefly the process of obtaining integrated 3D point clouds from plaster dental models. Section 3 introduces the segmentation techniques applied to 3D digital dental models. Section 4 shows the analysis and results of applying those segmentation techniques to point clouds of dental models. Finally, Section 5 presents the findings and discussion of the results of this work.

2 Dental Study Model

Experts in dental areas use diagnostic logs, which are kept in order to document the initial condition of the patient and complement the information gathered during clinical examination. These records are commonly divided in three categories: *dental models*, *photographs* and *radiographs* [10].

Dental models in dentistry are built using alginate cast. They are important for diagnosis and orthodontic treatment planning, as well as to detect anomalies of pose, size and shape of the teeth. Also, they are indispensable to assess the outcomes of treatment process [11] [12] [13].

2.1 Acquisition of point clouds of dental models

The Universidad Nacional de Colombia sede Manizales owns a 3D digitizer VIVID 9i Konica, Figure 1. This scanner produces range images which constitute a valuable source of information. Since range images cover the object's geometry from a specific point of view, several shots are needed in order to reconstruct a whole model without occlusions.

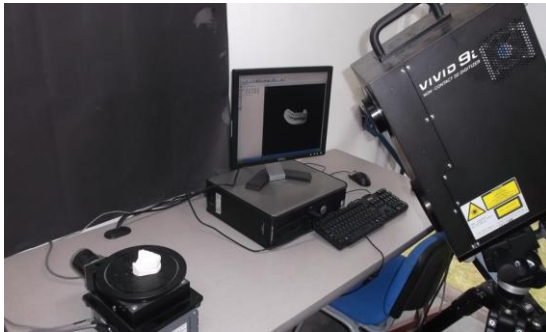


Figure 1: Konica Minolta VIVID 9i 3D digitizer.

The views acquired with the range scanner must be aligned up into a single coordinate space. This procedure is called registration, e.g. in Figure 2.

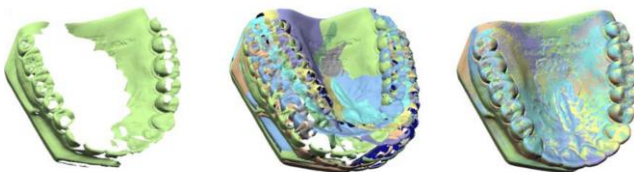


Figure 2: Registration.

Once the views have been registered, the integration process starts. The goal of integration is to generate a well-defined mesh or data set using the information coming from all the views (partial meshes or point sets) captured during scanning, Figure 3. Furthermore, this process seeks to eliminate redundant information in regions with little variation in the surface, and to fill small holes in the surface.

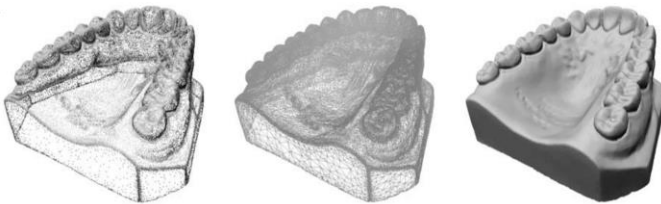


Figure 3: Integration.

In this study five dental models were used. Eight views were captured in order to construct each model, with the scanner tilted at 45 degrees, thus achieving good detail within the object of interest. This process was conducted by the research group in Perception and Intelligent Control (PCI). For further details please refer to the Master's final technical report Morantes [14].

3 Point Cloud Segmentation

In this section, a brief introduction on three different techniques for segmenting point clouds is given: RANdom SAMpling and Consensus (RANSAC) [15], Region Growing [16] and Maximum-Flow Minimum-Cut (Min-Cut) [17].

3.1 RANSAC

The iterative method of RANdom SAMpling and Consensus (RANSAC) was proposed by Fischer and Bolles. The technique aims to estimate the parameters of a mathematical model from a set of observed data using a method of hypothesis-testing. The algorithm is used as a geometric model based segmentation algorithm, due to its ability to automatically recognize shapes through the data (planes, cylinders, spheres and tori) despite the presence of noise.

The principle of the algorithm is as follows. If ϵ is the probability of choosing a sample that produces a poor estimate (outlier), then $1 - \epsilon$ is the probability of getting a good sample (inliers). This means that the probability of catching s good samples becomes $(1 - \epsilon)^s$. For k trials, the probability of failure becomes $(1 - (1 - \epsilon)^s)^k$. If ρ is the desired probability of success, then we have that:

$$1 - \rho = (1 - (1 - \epsilon)^s)^k \Rightarrow k = \frac{\log(1 - \rho)}{\log(1 - (1 - \epsilon)^s)} \quad (1)$$

3.2 Region growing segmentation

The Region Growing algorithm is based on the idea that some features in local data do not change greatly regarding those features measured for a given sample, called seed. Therefore, regions can be "grown" if neighboring data remain homogeneous given certain constraints, such as surface smoothness, curvature, etc.. In our work, these restrictions are based in local surface orientation and curvature, both approximated at each point in the cloud. The algorithm can be summarized as follows:

- Points are sorted according to their curvature.
- The point that has the minimum value of curvature is chosen as a seed and the region growth begins from that point (flat areas have less curvature).
- For each seed point, a list containing its closest neighbors is extracted:
 - The angle between the normal of each neighbor and the normal of the current point (seed) is compared: if the angle is less than a certain threshold, the point is added to the current region.
 - Subsequently, every neighbor is tested for the curvature value. If the neighbor's curvature is less than certain curvature threshold value then that point is turned into a new seed.
 - Current seeds are removed from the set of seeds, but remain marked as points belonging to the region. This avoids double checking points.

A region is said to be found once the algorithm runs out of seeds. According to Ushakov [16] the pseudocode is as follows.

Pseudocode: Region Growing

Inputs: Point cloud = $\{P\}$, Point normals = $\{N\}$, Points curvatures = $\{c\}$, Neighbour finding function $\Omega\{\cdot\}$, Curvature threshold c_{th} , Angle threshold θ_{th}

Initialize: Region list $R \leftarrow \emptyset$

Available points list $\{A\} \leftarrow \{1, \dots, |P|\}$

Algorithm:

While $\{A\}$ is not empty **do**

Current region $\{R_c\} \leftarrow \emptyset$

Current seeds $\{S_c\} \leftarrow \emptyset$

Point with minimum curvature in $\{A\} \rightarrow P_{\min}$

$\{S_c\} \leftarrow \{S_c\} \cup P_{\min}$

$\{R_c\} \leftarrow \{R_c\} \cup P_{\min}$

$\{S_c\} \leftarrow \{S_c\} \setminus P_{\min}$

For $i = 0$ to **Size** $\{S_c\}$ **do**

Find nearest neighbours of current seed point

$\{B_c\} \leftarrow \Omega(S_c\{i\})$

For $j = 0$ to **Size** $\{B_c\}$ **do**

Current neighbour point $P_j \leftarrow B_c\{j\}$

If $\{A\}$ contains P_j and

$\cos^{-1}\left(\left|\left(N\{S_c\{i\}\}, N\{S_c\{j\}\}\right)\right|\right) < \theta_{th}$ **then**

$\{R_c\} \leftarrow \{R_c\} \cup P_j$

$\{A\} \leftarrow \{A\} \setminus P_j$

If $c\{P_j\} < c_{th}$ **then**

$\{S_c\} \leftarrow \{S_c\} \cup P_{\min}$

end if

end if

end for

end for

Add current region to global segment list

$\{R\} \leftarrow \{R\} \cup \{R_c\}$

end while

Return $\{R\}$

3.3 Min-Cut

According to the results of Goloviskiy et al [17], Min-Cut is robust to noise and is very effective for segmentation in dense point clouds of outdoor urban scans but one drawback of Min-Cut is that it requires prior knowledge of the location of the objects to be segmented; also, two parameters (namely radius and sigma) should be set in order to control the resulting segmentation, where the main cues are distances and point densities, rather than colors and textures. The principle of the algorithm is as follows:

First of all the structure of the point cloud as a flow graph through the K closest neighbors, which are joined together using links with different weights. Additionally, points are joined to a source and a sink, used to represent the object and the background.

The first weight is given to all edges of the point cloud and is called *SmoothCost* (SC), computed using (2):

$$\text{SmoothCost} = e^{-\left(\frac{\text{dist}}{\sigma}\right)} \quad (2)$$

Where *dist* is the distance between the points and σ is a free input parameter that allows modifying the smoothing effect.

The next step of the algorithm is to establish the cost of the data. In this case it is necessary to make use of the source (t) and sink (s), where the source is related to the central point of the object to segment (given manually) and sink with any point belonging to background, where the following penalty calculated by equation (3).

$$\text{BackgroundPenalty} = \left(\frac{\text{distocenter}}{\text{radius}}\right) \quad (3)$$

Where *distocenter* is the expected distance to the center of the object in a horizontal plane (4).

$$\text{distocenter} = \sqrt{\left((x - \text{centerX})^2 + (y - \text{centerY})^2\right)} \quad (4)$$

And *radius* is an input parameter and can be considered as the range from the center of the object where the points belong to a region of interest by assigning a higher weight. On the other hand, *z* is not considered in (4); hence, the constraint can be geometrically interpreted as a cylinder from *z* to *-z*.

Once the graph is constructed, the segmentation is given by the minimum cut where the region of interest is extracted.

4 Results of segmentation techniques applied to 3D dental models

Exploration results obtained by applying different segmentation techniques described in Section 3 are shown in this section. Five 3D dental models were used in this analysis, Figure 10. Each cloud has approximately 50,000 points.

The tests were performed on all models; some important conclusions drawn from this analysis are mentioned. Finally a semi-automatic segmentation achievement is presented for 3D dental models.

4.1 Plane segmentation using RANSAC

As mentioned in Section 3.1, RANSAC allows automatic recognition of known forms of the data (planes, cylinders, spheres and tori). In this case a planar model was used, with a threshold of ± 5 mm; in Figure 4. We can see that threshold.

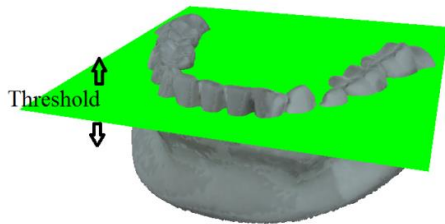


Figure 4: Planar segmentation by RANSAC with a threshold of ± 5 mm.

The result obtained applying RANSAC is shown Figure 5. The points in red color indicates the region extracted by RANSAC.

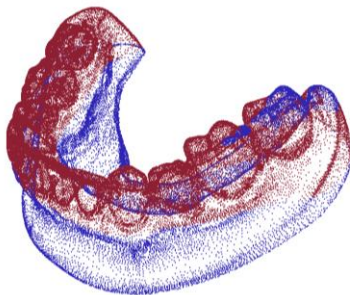


Figure 5: Dental model segmented using a planar model via RANSAC.

This method finds the highest concentration of points in a planar model based in a planar hypothesis. In this case the highest concentration of points is found on the teeth's surface.

4.2 Region growing segmentation of point clouds dental models.

The results obtained by applying the algorithm of region growing are shown in Figure 6 and Figure 7. The stop criteria are given comparing the points normal and then the curvature of the points. The input parameters are shown in Table 1.

Table 1, Input parameters for region growing segmentation

c_{th}	θ_{th}	N	$\Omega\{\cdot\}$	P
1	5	12	Tree	Point Cloud

In order to improve understanding and visualization, the point cloud of dental model is drawn with colors representing their curvature Figure 6.

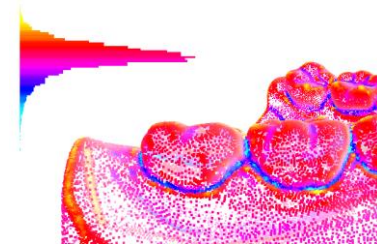


Figure 6: Example of curvature in point clouds of dental models.

For further information regarding neighbor search process and Kd-trees please refer to PCL API Documentation [18] or Flann library documentation [19].

The result obtained applying the Region growing segmentation technique in a 3D dental model is illustrated in Figure 7.

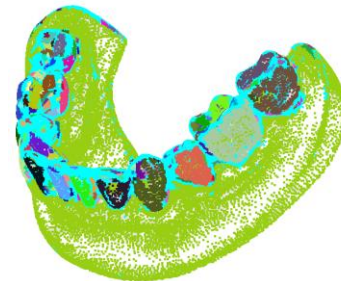


Figure 7: 3D dental model segmented using a region growing segmentation.

Figure 7. Shows some points colored in aqua blue, who are not targeted because of their high curvature or because the corresponding regions don't have the minimum size to belong to a region.

Clearly this method requires some refinement that allows merging some regions. However, this is an acceptable approximation usable as start point for further refinement.

4.3 Min-Cut segmentation

As mentioned in section 3.3, this algorithm requires human interaction to introduce the virtual node called source (t) to segment an object of interest. We can see a clear example in Figure 8. The input parameters are shown in Table 2.

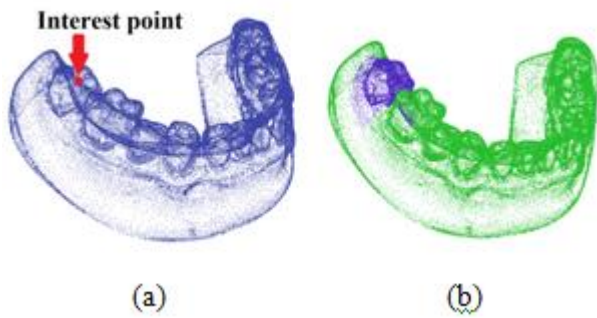


Figure 8: Segmentation applying the Min-Cut algorithm. (a) Selection of the interest point. (b) Segmentation of the selected tooth.

Min-Cut requires human interaction in order to set some important parameters such as the radius. For this reason, we designed a variant of the methodology used by Tamayo-Quintero et al, [20] to search the virtual nodes automatically – landmarks- i.e. a point in the centers of each object -in the ideal case-, in this case each tooth, by means of Normal Aligned Radial Features (NARF).

Table 2: Input parameters for Min-Cut segmentation.

Interest point (x,y,z)	σ	SmoothCost	radius
20,0,18	0.5	0.6	15

4.4 Semi-Automatic segmentation (Hybrid Technique)

In Figure 9. NARF detected interest points in the image automatically, given two parameters, the angular resolution and support size (For further details please refer to Steder [21]).

- The angular resolution (ag) mainly affects the size of the range image (deg/pixel).
- The support size (ss) is the diameter of the sphere that includes all the points used to calculate interest points.

As mentioned above, we designed a variant of the work in [20]. This methodology is based in the hybridization of three algorithms: region growing, NARF and Min-Cut.

- First of all, the input parameters are set according to Table 3.
- Next, gum is separated from teeth using the region growing method.
- Then, a set of landmarks is extracted using the NARF technique in the segmented teeth region.
- Subsequently, each landmark is used as source in order to apply the Min-Cut method iteratively.
- Finally, the segmentation is composed of the gum and a set of tooth regions.

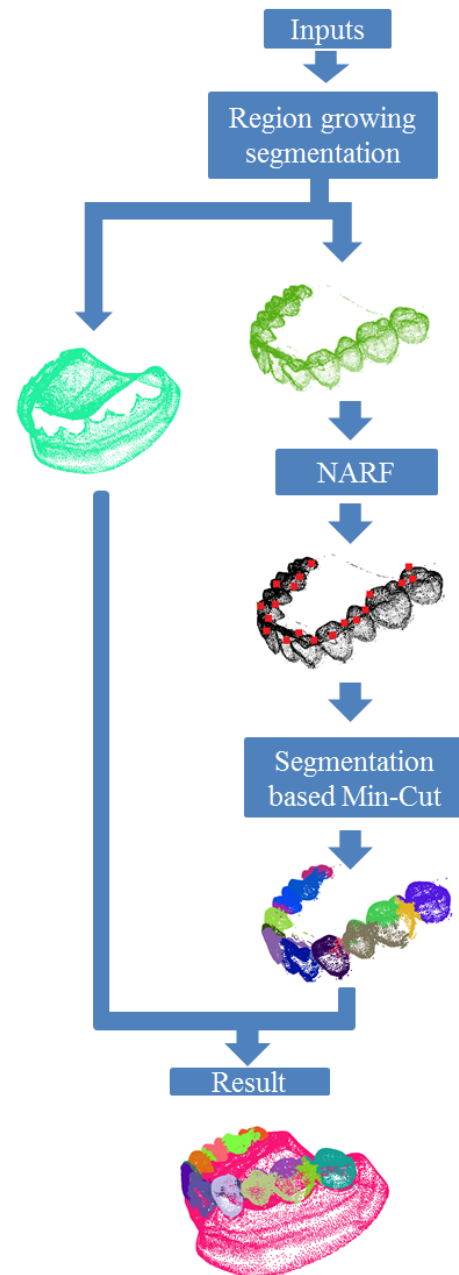


Figure 9: Semi-automatic segmentation proposal.

The methodology is illustrated in Figure 9, and the results obtained with it are shown in Figure 10.

Table 3, Input parameters used in the proposed methodology.

c_{th}	θ_{th}	σ	SC	radius	ag	ss	N
1	5	0.5	0.6	15	0.02	20	12


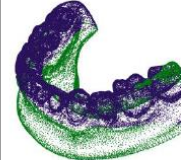

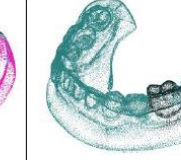

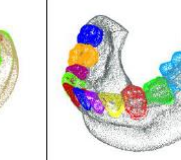
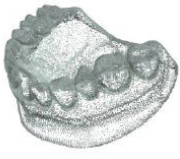
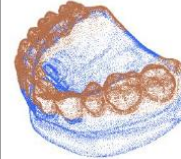
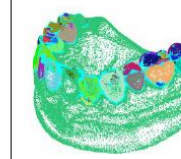
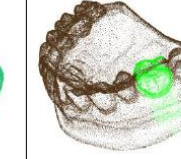
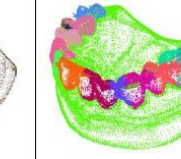
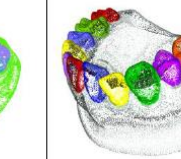


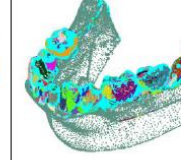
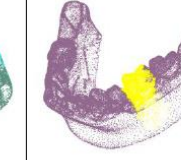




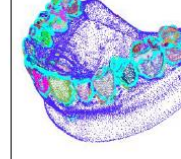
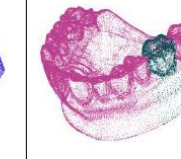
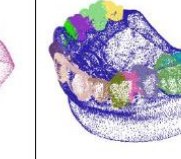
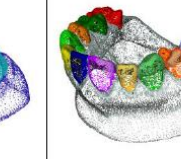

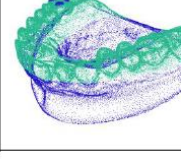
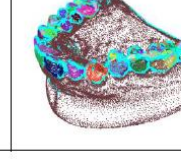
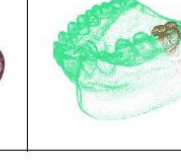
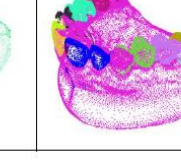
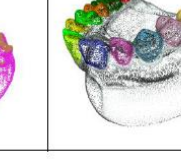
Dental Model	Plane Model (RANSAC)	Region Growing	Min-Cut	Hybrid Technique	Ground Truth
					
					
					
					
					

Figure 10, The results obtained by applying different segmentation techniques to five dental models.

5 Comments and discussions

Figure 10. Shows the result of applying segmentation methods to five dental casts. The first column contains the point cloud of dental models one to five. The second column shows the segmentation using RANSAC algorithm. There is a noticeable difference in result for model three regarding the other four, mainly due to teeth unevenness.

The third column presents the result of the region growing segmentation, where restrictions of smoothness given by the curvature and the angle between the normal are used. Results in this case are valuable and important, because the region of the gum and teeth are separated properly in every case (at first glance), but whose quality should be submitted to consideration and evaluation made by an expert.

The fourth column corresponds to Min-Cut based segmentation, which adequately separates the tooth but requiring human interaction in selecting a landmark for each tooth. As a consequence of this, the fifth column shows a methodology where landmarks are automatically selected through NARF, and then segmented using Min-Cut.

Finally, the sixth column shows the Ground truth, built manually using MeshLab.

6 Conclusion

In this paper we present an exploratory analysis of point cloud segmentation of dental models using well known segmentation techniques, where some achievements were obtained:

- Segmentation of the gum and teeth was carried out using the region growing algorithm based on the curvature and the angle between the normal.
- The potential for Min-Cut Algorithm for segmenting each tooth is evidenced.
- It is possible to apply the proposed methodology using NARF and Min-Cut, obtaining acceptable results for segmentation while reducing human interaction in setting internal parameters of the Min-Cut algorithm (landmark selection). Automatic radii selection remains an open issue.

7 References

- [1] Rusu, R.B. & Cousins, S., (2011). 3D is here: Point Cloud Library (PCL). *2011 IEEE International Conference on Robotics and Automation*, pp.1–4, available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5980567>.
- [2] FARO's 3D metrology solutions, Measurement Industries, available at: <http://www.faro.com/measurement-solutions>.
- [3] Motohashi, N. and T. Kuroda. A 3D computer-aided design system applied to diagnosis and treatment planning in orthodontics and orthognathic surgery. *Eur J Orthod* 1999. 21:263–274, available at: <http://ejo.oxfordjournals.org/content/21/3/263.full.pdf>.
- [4] Bell, a, Ayoub, a & F. Siebert, P., (2003). Assessment of the accuracy of a three-dimensional imaging system for archiving dental study models. *Journal of orthodontics*, 30(3), pp.219–23, available at: <http://www.ncbi.nlm.nih.gov/pubmed/14530419>.
- [5] Laurendeau, D., Guimond, & L. Poussart, D., (1991). A computer-vision technique for the acquisition and processing of 3-D profiles of dental imprints: an application in orthodontics. *IEEE transactions on medical imaging*, 10(3), pp.453–61, available at: <http://www.ncbi.nlm.nih.gov/pubmed/18222848>.
- [6] Mokhtari, M. & Laurendeau, D., (1994). Feature Detection on 3-D Images. , pp.287–296, available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=315842>
- [7] Chuah, J.H. Ong, S. H. & Kondo, T., (2001) “3d space analysis of dental models,” *Visualization, Display, an Image-Guided Procedures*, Proc. of SPIE, vol. 4319, pp. 564–573, 2, 26, available at: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=905781>
- [8] Heo, H. & Chaeb, O.-S., (2004). “Segmentation of tooth in ct images for the 3d reconstruction of teeth,” *Image Processing: Algorithms and Systems III*, Proc. of SPIE-IS&T Electronic Imaging, vol. 5298, pp.455–466, 2, 13, available at: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=837263>
- [9] Mok, V. W. Ong, S. Foong, K. W. & Kondo T., (2002). “Pose estimation of teeth through crown-shape matching,” *Medical Imaging 2002: Image Processing*, Proc. of SPIE, vol. 4684, pp. 955–964, available at: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=879917>
- [10] Hou, H.-M. Wong, R.W.-K. & Hagg, U. (2006). “The uses of orthodontic study models in diagnosis and treatment planning,” *Hong Kong Dental Journal*, vol. 3, no. 2, pp. 107–115, December 2006. 3, 14, available at: http://www.hkda.org/hkdj/V3/N2/v3N2_P107_DPI.pdf
- [11] Vellini-Ferreira F, (2002). *Ortodoncia: Diagnóstico y Planificación Clínica*, 1st ed. Editora Artes Médicas Ltda, 2002. xvii, 3, 5, 6
- [12] Williams, Acosta, F. Meneses, A. Morzán, E. Pastor, S. and Tomona, N. (1999) *Manual de procedimientos de laboratorio en ortodoncia*, 1st ed. Universidad Peruana Cayetano Heredia, 1999. 3
- [13] Rudge, S. J. (1982). “A computer program for the analysis of study models,” *European Journal of Orthodontics*, vol. 4, pp. 269–273, 1982. 9
- [14] Morantes, L. J. (2008) *Caracterización de piezas dentales a partir de modelos 3D*. Maestría tesis, Universidad Nacional de Colombia, Sede Manizales, available at: <http://www.bdigital.unal.edu.co/3377/#sthash.4QRtCEp7.dpuf>
- [15] Schnabel, R., Wahl, R. & Klein, R., 2007. Efficient RANSAC for Point-Cloud Shape Detection. *Computer Graphics Forum*, 26(2), pp.214–226, available at: <http://doi.wiley.com/10.1111/j.1467-8659.2007.01016.x>.
- [16] Point cloud Library (PCL), documentation, tutorial, available at: http://pointclouds.org/documentation/tutorials/region_growing_segmentation.php.
- [17] Golovinskiy, A. & Funkhouser, T., 2009. Min-cut based segmentation of point clouds. *2009 IEEE 12th International Conference on Computer Vision Workshops ICCV Workshops*, XXXIX-B3(September), pp.39–46, available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5457721>..
- [18] Point cloud Library (PCL), API documentation, available at: <http://docs.pointclouds.org/trunk/>.
- [19] Fast Library for Approximate Nearest Neighbors (FLANN), available at: <http://www.cs.ubc.ca/research/flann/>
- [20] Tamayo-Quintero, J.D. & Gómez-Mendoza, J.B., 2013. *Metodología para la Segmentación Semi-Automática de Nubes de Puntos 3D Empleando Min-Cut y NARF*. Encuentro Nacional de Investigación y Desarrollo (ENID 2013), available at: <http://www.enid.unal.edu.co/2013/memorias/>
- [21] Steder, B. et al., 2011. Point feature extraction on 3D range scans taking into account object boundaries. *2011 IEEE International Conference on Robotics and Automation*, pp.2601–2608. available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5980187>

Automatic Mass Segmentation Method in Mammograms Based on Improved VFC Snake Model

Xiangyu Lu, Yide Ma, Weiyang Xie, and Tongqing Li

School of Information Sci. Eng, Lanzhou University, Lanzhou, China

Abstract - Mammography analysis is an efficient way for the early detection of breast cancer. In this paper, we present an integrated method for mass auto-segmentation in breast. First of all, the local threshold method, Rough Set theory and morphological filter are used to remove the label and enhance the mammogram. Secondly we apply the Hough Transformation algorithm on the pre-processed image and locate the lesion as an approximate parametric circle which would be used as the initial contour of Snake model followed by. Finally, the mass boundary is accurately segmented by the improved VFC Snake. This approach is tested on DDSM and MIAS database and the detection rates are 91.47% and 85% respectively. The average area overlap ratio between our results and the ground truth on MIAS database reaches 90.4073%. The promising results indicate that our approach can provide some theoretical basis for computer-aided image detection system.

Keywords: Mammography; Early Breast Cancer Detection; Mass Segmentation; VFC Snake Model

1 Introduction

Breast cancer is one of the common malignant tumors and remains the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths in the world [1]. Mammography is a preferred method for early detection and also the most efficient and reliable tool for early prevention and diagnosis of breast cancer [2]. Mammograms are always with low contrast and the lesions are blurry and irregular, the shape and size of each mass or calcification are changeable, which cause the high misdiagnosis rate of breast cancer. For the past few years, computer-aid diagnosis has become the international research hot spot worldwide [3], which offers the doctors a reliable "second suggestion".

Best mass is an important symptom and its accurate segmentation is crucial to the treatment of breast cancer. Different algorithms of a system for detection of early lesion area in mammograms have been widely studied. Said Jai-Andaloussi et al. [4] proposed a bidimensional empirical mode decomposition method to segment the mass, while this method can extract the ROI only when the mass general region is determined. Tinxin Wan et al. [5] used variational method and global optimal active model to detect the mass,

unfortunately, the mass could not be accurately segmented out when the image contained interference regions. An integrated approach of contourlet transforms and phase portrait analysis etc. was applied to detect the architectural distortion by S.Anand et al. [6], which is not commonly adaptable due to its low detection rate however.

Novel image detection methods are appearing along with the development of technology. In recent years, the active contour model (Snake) [7] is widely used in image processing and computer vision, etc.[8,9], and among which VFC (Vector Field Convolution) Snake[10] model performs more excellent characteristics in segmentation of boundaries such as low dependence to initial contour, capability of convergence and superior noise robustness. While, it doesn't work well when we apply the typical VFC Snake method to extract the mass in mammograms, because of that the mass boundaries are always with low-contrast and appearing blurry in the whole image.

Considering the disadvantage mentioned above, we proposed an integrated approach for mass auto-segmentation in breast based on the improved VFC Snake model. The present method can detect the regions of masses in mammograms automatically and achieve promising results. This paper is organized as follows. In Section II, we present the methodology for mass localization and segmentation. Section III illustrates some experiments to verify the proposed method. Besides, the comparisons with typical VFC Snake model and the discussion also can be found in the Section III. Section IV gives the conclusions of this paper.

2 Methodology

The proposed methodology of breast mass segmentation can be schematically described in Fig.1. The method consists of four main processing steps: 1) Obtaining the mammogram images; 2) Mammogram pre-processing, to remove the label and enhance the image; 3) Mass localization: for determining the regions of interest and mass location parameters; 4) Mass accurate segmentation.

2.1 Mammogram database

The mammograms used in this work are taken from Digital Database for Screening Mammography (DDSM)[11] and Mammography Image Analysis Society (MIAS)[12]. The DDSM database consists of 2620 cases and each image was

about 3000×5000 pixels. The suspicious regions of each abnormality were provided with chain code data. The MIAS offered some corresponding information of lesion area such as type, location, severity, central coordinate and radius by experts and each image is 1024×1024 pixels.

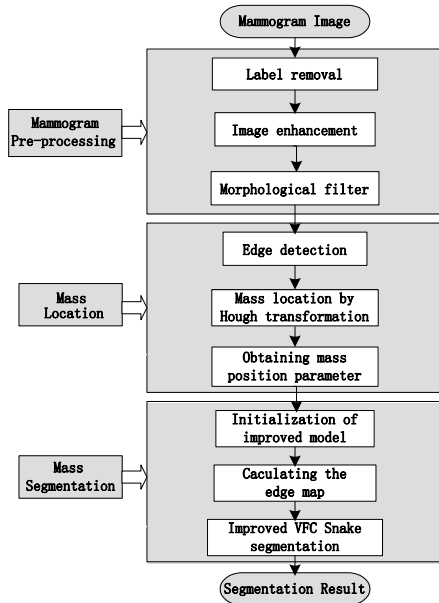


Fig.1. Flow chart of mass segmentation methodology

2.2 Mammogram pre-processing

2.2.1 Label removal

The mammogram image usually includes of breast region, pectoral muscle, background and label. In order to reduce the processing time and further study, the label should be removed. Here we use the standard local threshold method which has been proved to be a convenient method to remove the label.

2.2.2 Image enhancement

To extract the mass region accurately and eliminate noise, the Rough Set (RS) theory [13] is applied. Traditional gray level transformation is not working well since the gray values between mass region, pectoral muscle and gland are similar. RS theory which is used in reasoning from imprecise data is applied in information processing and artificial intelligence widely [14,15]. In this part we choose image gradient attribute C_1 and noise attribute C_2 as the condition attributes. According to the indiscernibility relation concept, the mammogram is divided into two sub-images:

$$R_{C_1}(i, j) = \{(i, j) \mid f(i, j) > P\} \quad (1)$$

$$R_{C_2}(s) = \bigcup_m \bigcup_n \{s_{mn} \mid \text{int} \mid \bar{s}_{mn} - \bar{s}_{m \pm 1, n \pm 1} \mid > Q\} \quad (2)$$

Where P is the gradient threshold, Q is the noise threshold, $f(i, j)$ is the gradient value calculated from the label-removed image, ' s ' denotes the sub-block. Considering s_{mn} as each pixel and examine its neighbors to decide whether it is noise or not. If it is, then eliminate the noise by replacing the pixel value with Q . The sub-images which need to be enhanced are defined as follows:

$$I_1 = R_{C_1}(i, j) - R_{C_2}(s) \quad (3)$$

$$I_2 = \bar{R}_{C_1}(i, j) - R_{C_2}(s) \quad (4)$$

Next, we enhance I_1 and I_2 respectively and get the final image by merging the sub-images: I_2 is enhanced by histogram equalization method and I_1 is transformed below:

$$g(i, j) = \rho \cdot I_1(i, j)^\gamma \quad (5)$$

Here we set $\rho = \gamma = 1.5$. After enhancement, the boundary contrast between the mass and surrounding tissue becomes more obvious.

2.2.3 Morphological filter

Then we amend the enhanced image using morphological filter. The pre-processed image is shown in Fig.2.



Fig.2. Image (MIAS) before and after enhancement and pre-processed result

2.3 ROI extraction and location

2.3.1 Edge extraction

The pre-processed image is composed of pectoral muscle and mass region. Before removing the pectoral muscle, the edge detection operator is used to extract the edge first.

2.3.2 Hough Transform detection

From Fig.3, the edge of pectoral muscle appears as a triangle while mass edge usually appears as an ellipse or circle, thus we can obtain the approximate edge of mass by performing the Linear Hough Transform[16] (LHT) and Circular Hough Transform (CHT) on the extracted edge image, which respectively defined as formulas (6) and (7). Here Ω_{pectoral} is the pectoral muscle detected by LHT on edge image, Ω_{mass} is the initially segmented mass obtained by CHT on the region $\Omega_{\text{edge-pectoral}}$ where Ω_{pectoral} is removed. (a_i, b_i) is the center position of the circle and R is the radius. The location results are shown in Fig.3.

$$\Omega_{edge} \rightarrow \Omega_{pectoral} : (\rho, \theta) = LHT(x_i, y_j) \quad (6)$$

$$\Omega_{(edge-pectoral)} \rightarrow \Omega_{mass} : \quad (7)$$

$$((a_i, b_i), R) = CHT((x_i, y_i), r_{min}, r_{max})$$

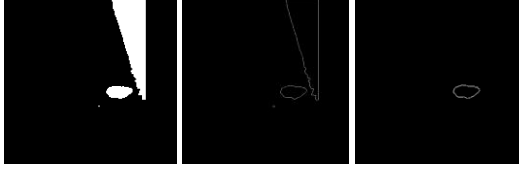


Fig.3. Edge extraction and mass location results(1024×1024 pixels)

2.3.3 Mass location parameter

For a further work, we obtain the mass position coordinate by defining a circle, whose center corresponds to the center of extracted mass edge, and the radius of which is a middle value of the distance from boundary pixel to the center position, indicating as: (cx, cy) and r . For Fig.3 the parameter is: $(682, 586)$, $r=46$. The parametric circle could be used as the initial contour of deformable model followed by.

2.4 Mass segmentation

The results of mass location are rough and exist certain gap with the actual boundary, we utilize the Snake model to perform an accurate segmentation further.

2.4.1 Typical VFC Snake model

The Snake model defined a parametric curve guided by external forces and internal forces that pull it toward the edge of ROI until the energy function achieves the minimum. The curve $v(s)$ and the minimizing energy function $E(v)$ forms are:

$$v(s) = [x(s), y(s)]^T, s \in [0, 1] \quad (8)$$

$$E(v) = \int_0^1 (E_{int}(v(s)) + E_{ext}(v(s))) ds \quad (9)$$

$$E_{int}(v(s)) = \frac{1}{2} [\alpha |v'(s)|^2 + \beta |v''(s)|^2] \quad (10)$$

Where E_{int} is the internal energy decided by the curve, α and β control the continuity and smoothness respectively. E_{ext} is the external energy decided by the image information. The Vector field convolution Snake is an active contour model whose external force is VFC field. First defined a vector field kernel:

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = [\mathbf{u}_k(\mathbf{x}, \mathbf{y}), \mathbf{v}_k(\mathbf{x}, \mathbf{y})] = m(x, y)\mathbf{n}(\mathbf{x}, \mathbf{y}) \quad (11)$$

Where $m(\mathbf{x}, \mathbf{y})$ is the vector magnitude and $\mathbf{n}(\mathbf{x}, \mathbf{y})$ is the unit vector pointing to the kernel origin $(0, 0)$.

$$m(x, y) = (r + \varepsilon)^{-\gamma} \quad (12)$$

$$\mathbf{n}(\mathbf{x}, \mathbf{y}) = [-\mathbf{x}/r, -\mathbf{y}/r] \quad (13)$$

Here $r = (x^2 + y^2)^{1/2}$. The external force is calculate by convoluting the vector field kernel $\mathbf{k}(\mathbf{x}, \mathbf{y})$ and the edge map $f(\mathbf{x}, \mathbf{y})$, defined as:

$$f_{vfc} = f(x, y) * \mathbf{k}(\mathbf{x}, \mathbf{y})$$

$$= [f(x, y) * \mathbf{u}_k(\mathbf{x}, \mathbf{y}), f(x, y) * \mathbf{v}_k(\mathbf{x}, \mathbf{y})] \quad (14)$$

2.4.2 Improved VFC Snake model

Firstly, we use the typical VFC Snake model to detect the lesion area, however, there are obvious distortions in the segmented results. By analyzing the force field of the model we observe that the distribution of which is disordered (Fig.4(b)(c)), and this causes the misleading of the active contour to the interference tissues rather than the real mass boundaries. Thus, the main idea of our proposed segmentation algorithm is to improve the force field by defining a new and clear edge map $f(\mathbf{x}, \mathbf{y})$. First, the gradient value of each pixel is improved using RS theory method by setting a gradient threshold to judge whether the pixel belongs to a potential boundary or not. If it is, then these pixels should be enhanced. Next, the edge map $f(\mathbf{x}, \mathbf{y})$ is calculated by performing the Canny operator [17] which is considered to have more excellent features like stability and accuracy compared to the gradient operator. As shown in Fig.4(d)(e), it is obvious that the improved force field distribution turns to be more regular, which appears much evenly near the boundaries and the areas mixed in normal glands.

Besides, the performance of typical VFC Snake model depends on the position of initial contour around the mass, which usually is a circle formed by the coordinate position and radius. Here we set the initial contour by using the parametric circle obtained in the mass location part.

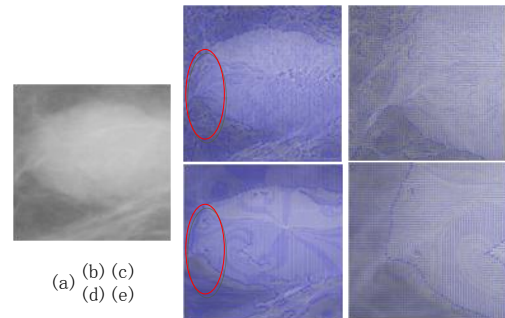


Fig.4. Force field of typical and improved VFC Snake and the enlarged images

3 Experiment results and discussion

Our proposed method was tested on 400 mammograms with abnormal breast regions from DDSM and MIAS database. The experimental results are shown in Fig.5 and Fig.6 respectively. We can observe that our approach can achieve much better results in comparison with the typical VFC Snake method. During this test, we set the parameter of VFC Snake model as: $\alpha = 0.5$, $\beta = 0.2$, and the iteration is about 30. To fully explain the superiority of our proposed method in visual, the illustrated mammograms consist of craniocaudal (CC) and mediolateral oblique (MLO) view, whose severity includes benign and malignant, and each case contains only one abnormality.

3.1 Experiments results

The suspicious regions of full raw images from DDSM were given by chain code data as ground truth. This database also provided thumbnail images for visual browsing of each case as shown in Fig.5(a)-(d). The severity of these selected images is benign and the abnormal regions are marked as red curves. Here we utilize the parametric circle obtained from mass location procedure as the initialization of VFC Snake model and segment the lesion area. Results gained by our method are shown in Fig.5(a-2)-(d-2), labels have been removed already. From the enlarged images, we can clearly see that the contours (blue curves) converge precisely to the real boundaries in all cases. For comparison, Fig.5(a-3)-(d-3) state the same cases detected by the typical VFC Snake model, unfortunately, these results are seriously influenced by the blurry tissues and can't deform to the objects completely, even the initial contours are very close to the actual boundaries. Thus we can say that the improved model is in lower dependence on the initial contour and with stronger capability of convergence than the typical method, that is, our approach performs much better when we segment the masses in mammograms from DDSM database.

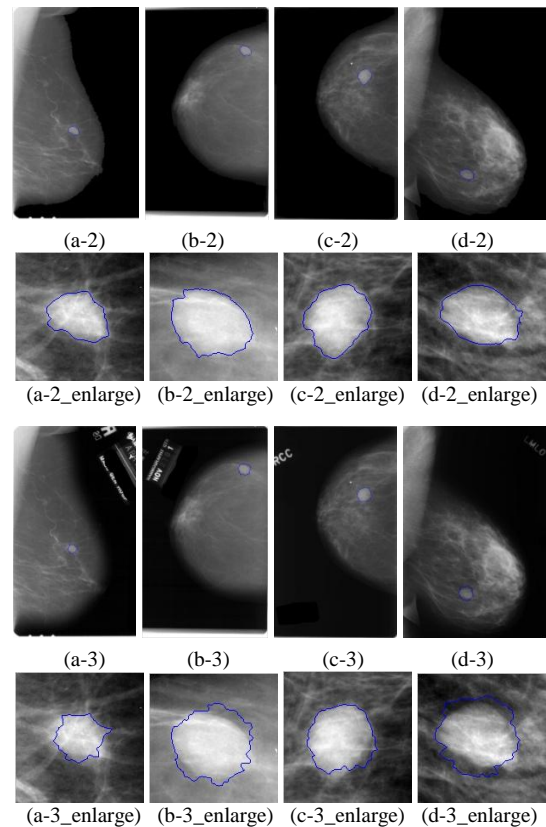
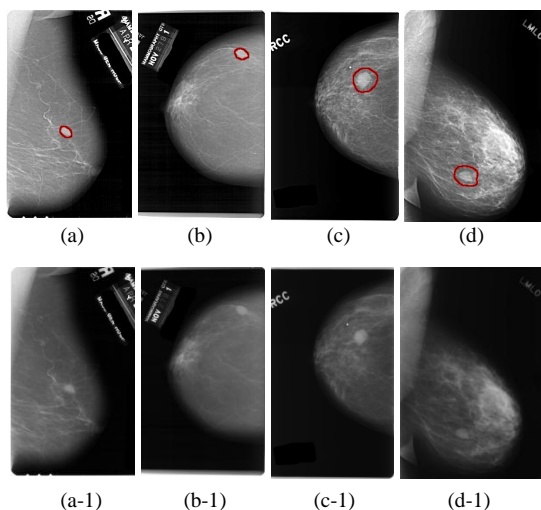


Fig.5. Experimental results of mammograms from DDSM database: (a)-(d)The "thumbnails".(a-1)-(d-1) Full raw images .(a-2)-(d-2)Results by our proposed method.(a-3)-(d-3)Results by typical VFC Snake model.

The MIAS database has offered the central coordinate and radius of each abnormal region forming as green circles showing in Fig.6((a-3)-(d-3)), the red curves are the ground truth segmented manually. Here we also initialize the VFC Snake model using the parametric circle obtained by mass location. From Fig.6((a-2)-(d-2)), we observe that the results (blue curve) segmented by the typical method exist serious distortions and the contours can hardly converge to the real boundaries. Compared with the typical model, our proposed method can completely remove the labels or interference and achieve more robust and accurate results. As we can see, the curves are much more close to the ground truth and precisely tend to the object even in blurry regions.

From the enlarged results of our method, we find the margin of the last image is rough and the other ones are smooth, that is because the severity between the last lesion and the rest are different, the last lesion is malignant while others are benign. Our results objectively reflect the pathology characteristics of actual masses to some extent that the malignant masses are always with burrs. This performance is somewhat benefit to the early diagnosis of breast cancer. Therefore, for a CAD system, we are able to extract the features of our detected results and determine the severity of abnormalities for a further work to give a considerable "second suggestion" to the clinician.

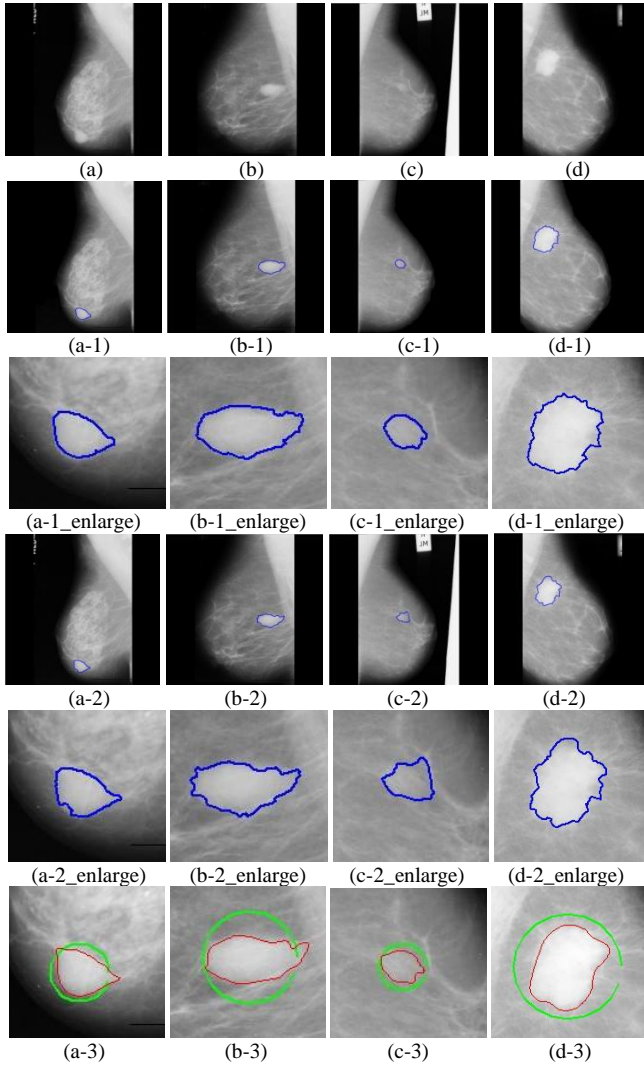


Fig. 6. Experimental results of mammograms from MIAS database: (a)-(d)Original images.(a-1)-(d-1)Results by the proposed method. (a-2)-(d-2)Results by typical VFC Snake model. (a-3)-(d-3)The ground truth.

3.2 Algorithm performance analysis

We test the proposed method on the DDSM and MIAS database respectively and evaluate the performance from two aspects.

3.2.1 Detection rate

Our evaluation principle is that the auto-segmented region by the proposed method is completely within the criterion region by the experts. In the case of DDSM database, the criterion region is the outline formed by chain code data, and for MIAS database the criterion region is the circle formed by the center coordinates and the radius. The detection rates of masses for each database are shown in Table I. As we can see, 362 images are successfully extracted in total, the average detection rate is 90.5% and even reaches up to 91.47% for the DDSM images. While it is lower for the MIAS images, because of that the lesions in dense breast images of MIAS are always embedded in the gland and we can hardly

obtain the mass position by location or edge map by edge detection operator for the deformable model.

TABLE I. MASS DETECTION RATE BY THE PROPOSED METHOD

Database	Tested images	Detected images	Non-detected images	Detection rate(%)
DDSM	340	311	29	91.47
MIAS	60	51	9	85.0
Total	400	362	38	90.5

3.2.2 Segmentation accuracy

To further explain the accuracy of our algorithm, we introduce another evaluation method. The area overlap ratio criteria is the most common evaluation criteria in medical images, which is the ratio of overlapped area between the segmented region of VFC Snake method and the criterion region of ground truth segmented manually. The performance of the proposed method and the typical VFC Snake method are tested on the successfully detected mammograms of MIAS by the equation below:

$$O = \frac{S_{L \cap T}}{S_{L \cup T}}$$

Where L is the area segmented by VFC Snake model, T is the area of ground truth. $S_{L \cap T}$ and $S_{L \cup T}$ are the intersection area and union set area of the two regions respectively. The average area overlap ratio and the variance of the segmentation results are shown in Table II. We can see that the average area overlap ratio of improved method is much higher than the typical method, and the variance is much lower, that is to say, our auto-segmented results are generally much more close to the ground truth. It is proved that our approach indeed performs much more excellent results compared with the typical method.

TABLE II. AREA OVERLAP RETIO OF DIFFERENT METHOD

Method	Mean (%)	Variance (%)
Typical method	76.0151	11.5249
Improved method	90.4073	2.1556

4 Conclusions

In this work, we present an effective integrated approach based on the improved VFC Snake model for mass automatic segmentation in mammograms which with low contrast and blurry boundaries. First of all, the local threshold method, RS theory and morphological filter are applied to pre-process the mammograms to remove the labels and enhance the whole image. Then we use the LHT and CHT algorithms to locate the massive lesions and the position of which parametrically indicated as an approximate circle. The mass segmentation stage uses the parametric circle to initialize the deformable

method which is defined by improving the force field of typical VFC Snake model and extract the mass boundary accurately. The proposed approach is tested on DDSM and MIAS database respectively and the results show that our algorithm achieves a higher detection rate and superior segmentation accuracy compared with the typical VFC Snake model. In conclusion, the improved approach can not only locate and segment the mass automatically, but also in lower dependence on the initial active contour and with stronger capability of convergence. Besides, this algorithm is robust to the interference of blurry areas and tissue and able to converge precisely to the object. What's more, the results conform to the pathology characteristics of actual masses to some extent and benefit to early detection of breast cancer. That is to say, the proposed approach can provide some important basis to improve the CAD system.

In future work, we would like to classify the breast masses to benign and malignant based on the auto-segmented results of this paper.

5 Acknowledgment

This work is jointly supported by the National Natural Science Foundation of China (Grant No.61175012), Science Foundation of Gansu Province of China (Grant No.1208RJZA265), Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No.20110211110026), and the Fundamental Research Funds for the Central Universities of China (Grant No.lzujbky-2013-k06).

6 References

- [1] Jemal, Ahmedin, et al. "Global cancer statistics." *CA: a cancer journal for clinicians*, vol.61, no.2, pp.69-90, 2011
- [2] Xu Guangzhong, Li Kai, Feng Guosheng. "Comparison of three imaging methods in the early diagnosis of breast cancer." *Journal of Capital Medical University*, vol.30,no.3, pp.293-297, 2009
- [3] Ouyang Cheng, Ding Hui, Wang Guangzhi. "Segmentation of masses in mammograms." *Beijing Biomedical Engineering*, 2007
- [4] Jai-Andaloussi, Said, et al. "Mass segmentation in mammograms by using Bidimensional Empirical Mode Decomposition BEMD." *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. IEEE*, pp.5441-5444, 2013.
- [5] Wan, Jinxin, et al. "Mammographic Mass Segmentation Using Variational Method." *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 5th International Conference on. Vol. 2. IEEE*, 2013.
- [6] Anand, S., and R. A. V. Rathna. "Detection of architectural distortion in mammogram images using contourlet transform." *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on. IEEE*, pp.177-180, 2013.
- [7] Kass, Michael, Andrew Witkin, and Demetri Terzopoulos. "Snakes: Active contour models." *International journal of computer vision. Vol.1, no.4,pp.321-331,1988*
- [8] Mouelhi, Aymen, Mounir Sayadi, and Farhat Fnaiech. "A supervised segmentation scheme based on multilayer neural network and color active contour model for breast cancer nuclei detection." *Electrical Engineering and Software Applications (ICEESA), 2013 International Conference on. IEEE*, pp.1-6, 2013.
- [9] Miao Guo, Zhaobin Wang, Yide Ma, et al. "Review of parametric active contour models in image processing." *Journal of Convergence Information Technology. Vol.8, 2013.*
- [10] Li, Bing, and Scott T. Acton. "Active contour external force using vector field convolution for image segmentation." *IEEE Trans.Imag.Proc.*, vol.16, no.8,pp.2096-2106,2007.
- [11] Bowyer K, Kopans D, et al. "The digital database for screening mammography." *Third international workshop on digital mammography*,vol.58, 1996.
- [12] Suckling, John, et al. "The mammographic image analysis society digital mammogram database." 1994.
- [13] Pawlak, Zdzisław. "Rough set approach to knowledge-based decision support." *European journal of operational research*, vol. 99, no.1, pp.48-57, 1997.
- [14] Abraham, A., et al. "Rough sets and near sets in medical imaging: a review." *Information Technology in Biomedicine, IEEE Transactions on*, vol.13, no.6, pp.955-968, 2009.
- [15] Liu, Ying-Jie, et al. "Rough sets theory and its applications in image processing." *Jisuanji Yingyong Yanjiu/ Application Research of Computers*,vol.24,no.4,pp.176-178,2007
- [16] Hough, Paul VC. "Method and means for recognizing complex patterns." *U.S. Patent No. 3,069,654. Dec. 1962.*
- [17] Canny, John. "A computational approach to edge detection." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* vol.6, pp.679-698, 1986.

Correction of Intensity Nonuniformity in Mammographic Images

S. Yazdani¹, R. Yusof², A. Karimian³, A. Hematian⁴

^{1,2}Centre for Artificial Intelligence & Robotics, University Teknologi Malaysia, Kuala Lumpur, Malaysia

³Department of Biomedical Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran

⁴Dept. of Advanced Informatics School (AIS), Universiti Teknologi Malaysia (UTM), Kuala Lumpur, Malaysia

Abstract- *Breast cancer is one of the most prevalent cancers among women[1]. Mammography, as one of the primary studies, is used for diagnosis of breast disease. In addition MR images can depict most of the significant changes of breast during the time. For the first step of breast disease detection, the density measurement of the breast on MR Images may provide very useful information. MR images have some instinctive limitations like the strongly dependence of contrast upon the way the image is acquired, intensity inhomogeneities (Bias Field), etc. For these reasons, an effective normalization on breast MR Images is very important issue for detecting breast disease signs. The first important step for quantitative analysis of breast density on MRI is Bias Field Correction. In this study, N3 algorithm is used for correcting of field inhomogeneity in MR images. We used T1-weighted images, using a 1.5 T MRI scanner. The results demonstrate effectiveness and efficiency of the proposed method.*

Keywords: Bias field correction, Breast disease diagnosis, MR Images, N3 method.

1 Introduction

Breast cancer is the second factor for death among women around the world [1]. Mammography is considered as one of the primary studies on tumor diagnosis and breast disease. It is not recommended to have frequent MRIs for high-risk women because of radiation. Scientifics are looking for the best way for quantifying breast densities of MR images in healthy women to assessment of healthy breast composition. They are trying to find the link between breast cancer and breast composition. In addition these assessments can help future researches to estimate potential risk of breast cancer. Our goal is normalizing MR images for quantifying density of the breast accurately, because this factor is an important marker in diagnosis of breast cancer. However, MR images has some limitations, such as: They sometimes are in low contrast, the dependence of MR image quality upon the condition the image is acquired, ideal image situation is never realized practically, Bias field, etc.

One of the main problems of MR Images is bias field[2]. Bias field has been a challenging problem in MR Image segmentation. It is a smooth and low-frequency signal that corrupts MRI images specially those produced via old MRI

machines. Bias field is attributed to eddy current, poor radio frequency coil uniformity and patient anatomy[3]. Bias field eliminate high frequency contents of MRI image like contours and edges, blurs the images, changes the intensity of image pixels, as a result, same tissues have different gray level distribution in the image. In order to decrease the aforementioned restriction, research teams throughout the world have conducted some studies on bias field correction in mammographic images [4];[5].

According their studies there are two main methods for bias field correction: prospective and retrospective methods. The prospective methods try to solve this problem in the process of acquisition by using special hardware. This approaches can only delete inhomogeneities due to hardware imperfections.

Retrospective approaches have been more developed; they are classified to two groups: first group use the segmentation based methods for computing the bias field and the second one work directly on data.

Although the segmentation based approaches such as (EM) expectation maximization algorithm[6], FCM-based methods[7](ML) maximum likelihood[8]; [9], MAP-based approaches [10]; [11], have obtained suitable result but they have some disadvantages such as: they work solely on intensity of image and they are able only to estimate and correct low amplitude intensity inhomogeneities[12].

The methods which work directly on MRI image data such as SPM99[13]. This method has a problem of entropy minimization. Another method in this category is N3(nonparametric intensity nonuniformity normalization) [3], that we used it in this paper for bias field correction. N3 method was determined to be the best method on the recent studies[3].

After digitalizing the images N3 method was used for bias field estimation and correction. The input of proposed method includes different percentages of intensity inhomogeneities while the output consists of bias field corrected images with very high quality which indicate breast regions more precisely.

2 Methodology

2.1 N3 Algorithm

Bias correction is one of the important stages in MRI processing. Nonparametric intensity nonuniformity normalization (N3) was proposed by Sled [3], for solving the problem of artifacts in MRI images. It is an iterative method which estimates multiplicative bias field and true tissue intensity distribution. This method is an intensity model based or histogram base approach. Unlike some other methods such as EM, N3 does not rely on tissue classification.

One of the main advantages of nonparametric methods is that doesn't make any assumptions about the patient anatomy. In addition it is fully automatic, accurate and robust method. In this paper the N3 algorithm is used in mammogram images.

2.2 How the N3 algorithm work?

The following equation is the basis of the N3 method[14]:

$$v(\mathbf{x}) = u(\mathbf{x})f(\mathbf{x}) + n(\mathbf{x}) \quad (1)$$

In which:

f : An unknown bias field

v : Measured signal

x : Location

u : True signal emitted by the tissue

n : The noise which

assumed to be independent of u

The main stage for correcting bias field is estimating its distribution (f). The mixture of multiplicative and additive model makes this stage difficult. Consider a case without noise in which u and f are independent distributed random variables. Instead of v, u, f we deal with $\log v, \log u, \log f$, then the formation model becomes additive:

$$\hat{v}(\mathbf{x}) = \hat{u}(\mathbf{x}) + \hat{f}(\mathbf{x}), \quad (2)$$

Where the $\hat{u}(x) = \log[u(x)]$

The multiplication corrupts the field and a division can undo the corruption. In the frequency domain, multiplications and divisions convert to convolutions and deconvolutions as follow:

$$V(\hat{v}) = F(\hat{v}) \times U(\hat{u}) = F(\hat{v} - \hat{u})U(\hat{u}) \quad (3)$$

In which $V, U,$ and F are probability densities. After this stage the uniformity distribution (F) (The main stage for

correcting bias field) is modeled and viewed as blurring intensity distribution U .

2.3 Correction step

Since the bias field degrades the high frequency components of the MR image, correcting intensity nonuniformity is the restoration of frequency content of U . Correction method for bias field is finding the slowly varying, smooth and multiplicative field which maximizes the frequency content of U (sharpened distribution of V).

The sharpening process is performed by using parameters Z and FWHM (full width at half maximum). The estimation of U is done as follows:

$$\hat{G} = \frac{\tilde{F}^*}{\tilde{F}^2 + Z^2} \quad (4)$$

and

$$\hat{U} = \hat{G}\hat{V} \quad (5)$$

In which:

$\sim F$ = the Fourier transform of F

F^* = complex conjugate

Z = a constant term to limit the magnitude of $\sim G$

This step is used for field estimation in the next stage.

2.4 Field estimation

The estimated value of \hat{u} given a measurement v_j is defined as follows:

$$E(\hat{u}) = \frac{\int_{-\infty}^{\infty} \hat{u} F(\hat{v} - \hat{u}) U(\hat{u}) d(\hat{u})}{\int_{-\infty}^{\infty} F(\hat{u}) U(\hat{u}) d(\hat{u})} \quad (6)$$

By using the estimation of \hat{u} the estimation of f can be found as follows:

$$\hat{f}_e(\hat{v}) = \hat{v} - E[\hat{u}|\hat{v}] \quad (7)$$

The difference between V and the computed expectation of real signal, given the measured signal.

In addition to processing stages were described above, there are some steps for practical implementation of N3 algorithm.

- Determining foreground by a simple threshold.
- Estimating V distribution by using an equal-size bins histogram and Parzen window[15].

$$V(\hat{v}_j) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h} \varphi \left[\frac{\hat{v}_j - \hat{v}(\mathbf{x}_i)}{h} \right] \quad (8)$$

$$\varphi(s) = \begin{cases} 1 - |s| & |s| < 1 \\ 0, & \text{elsewhere} \end{cases} \quad (9)$$

In which

h :The distance between v_j .

x_i :Location

N :Set of measurements $v(x_i)$

v_j : Centers of the bins

- Smoothing the bias field by using the B-spline technique [16].For doing this stage the MRI data should resample into subsamples (coarser resolution). This step is because of smoothing the bias field at full resolution is computationally difficult.

The smoothing stage is a challenging stage and the (10) of smoothing, effects on bias field correction performance. The proposed approach to smoothing is approximating data by using linear combination of smooth basis functions. B

$$f_s(\hat{v}) = S\{f_e(\hat{v})\}$$

spline is a suitable basis which is compactly supported spline. In comparison with conventional filtering approaches the proposed technique is superior regarding to missing data. Filtering methods are not suitable for this step because of boundary effects which degraded overall performance substantially. For example low pass filters may cause some problems in large breasts. The boundary of the chest wall of fatty tissues may be smoothed out to be close to outside background . Smoothing filter methods, assume that intensity inhomogeneity is low frequency and the other components have higher frequency, which is usually wrong specially for dense breast cases, and the bias field correction would lead to erroneous for diagnosis of fat and fibroglandular tissue [13].

- Resampling to original resolution and using it for correcting original volume.
- Terminating the iterations ,as follow:

$$e = \frac{\sigma\{r_n\}}{\mu\{r_n\}}, \quad n = 1 \dots N$$

r_n : Ratio between subsequent field estimates at the location

σ : Standard deviation

μ : Mean

When e falls below 0.001 the iteration is stopped. In this paper breast MR images including variety percentages of bias field were used.

By using mentioned method the bias field is removed and the specialist can assess and quantify the density of the breast accurately, and also the bias field corrected images are ready for another processing such as segmentation.

More importantly,segmentation and bias field estimation are mutually influenced by each other and the performance of MRI segmentation can be degrade significantly by presence of bias field.If the bias field is corrected, the segmentation would be more powerful and it helps to specialist to have an accurate assessment .

3 Experimental result

This paper demonstrated recent progress on MR image bias field correction. As mentioned above there are different methods which are popular to bias field correction such as low pass filtering and statistical methods. Low pass filtering techniques are fast, easy to code in addition they can also be adaptive to image data .One of the main disadvantages of low pass filtering method is assuming the bias field is low frequency signal .This methods also assume that the other image component have higher frequency ,which is usually wrong for some cases. In addition they tend to corrupt low frequency components in tissue.

Statistical techniques are easy to integrate with other knowledge such as registration ,segmentation or some image feature but one of the disadvantages of these methods is, they often have relied on Gaussian distribution for modeling the intensity distribution of tissues, but experimental results show that intensity distribution of tissue do not indicate a Gaussian mixture exactly .This method fails to present some tissues[17].

Overall, the N3 method is superior to the other methods in robustness and high performance point of view[18].

In this part, we show the efficacy of N3 algorithm for clinical breast MRIs. All of the images were acquired by 1.5T clinical MR scanner in ["Isfahan radiology Medical Center"].Fig.1 (a) demonstrates a breast MR image corrupted by bias field. Fig. 1 presents the estimated bias field by using mentioned algorithm and the corrected image by using N3 algorithm.

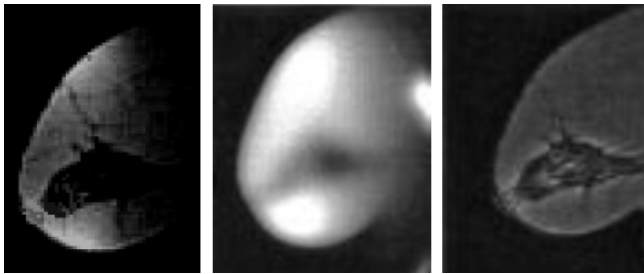


Figure 1: Sample MR image with severe bias field, left to right: Original MR Image , estimated bias field and corrected image

Mentioned algorithm estimated bias field, corrected it and improved the image contrast and quality dramatically. Fig.2 indicates other examples of breast MRI. Fig.2 (a) is the original MR images and Fig.2 (b) is bias field corrected images respectively.

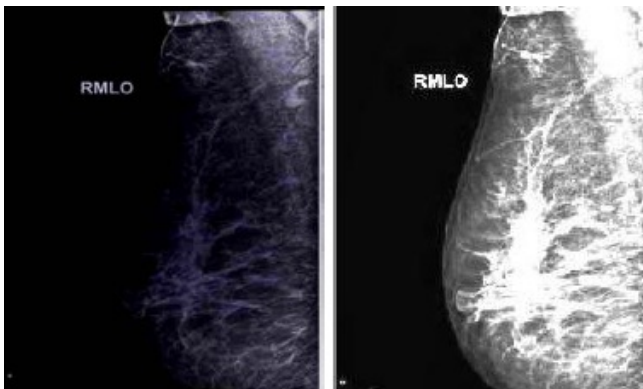


Figure 2: Original MR image with sever bias field(a),corrected image(b)

We can see that the bias field in our clinical database is 30% 40% and the algorithm has removed the bias field successfully.

Despite of this fact that in figures 1(a)and 2(a), mammographic digitalized images have different percentages of bias field. These images are blur because of bias field existence and breast compositions are not clearly shown. In the images which are normalized through mentioned algorithm , these regions are presented clearly in figures 1(c)and2(b). In these images the dynamic range is increased and the fatty tissues became brighter than before and they can be separated from fibroglandular tissues.

4 Conclusion

The bias field is the challenging problem of MR images. It changes the intensity values of pixels in MRI image and

corrupts these images. The correction of this problem is necessary for subsequent computerized quantitative analysis. The high importance of bias field correction, motivated us to use a fast, reliable, and robust algorithm to solve this problem.

In this paper, an effective, robust and accurate method is used for bias field correction in breast MR Images. N3 method is an iterative algorithm, and does not need any model assumption. This method does not rely on any prior knowledge of pathological data as a result can be applied at early stage and it is a substantial advantage in automated data analysis .Breast MRI image as an input enters into the proposed package and after using several processing methods, the output is an image which indicates breast compositions and density measurement of the breast. The efficacy of the algorithm is presented on clinical breast MRIs and the results show the potential of method to extract useful information for breast disease detection.

5 Acknowledgment

The authors would like to thanks Ministry of Education of Malaysia (MoE), University Teknologi Malaysia (UTM) and Research Management Centre of UTM for providing financial support under the GUP Research Grant No. 04H40, titled "Dimension reduction and Data Clustering for High Dimensional and Large Datasets".

6 References

- [1] Avril, N., et al., *Utility of PET in breast cancer*. Clinical Positron Imaging, 1999. **2**(5): p. 261-271.
- [2] Roy, S., et al. *Intensity inhomogeneity correction of magnetic resonance images using patches*. 2011.
- [3] Sled, J.G., A.P. Zijdenbos, and A.C. Evans, *A nonparametric method for automatic correction of intensity nonuniformity in MRI data*. Medical Imaging, IEEE Transactions on, 1998. **17**(1): p. 87-97.
- [4] Kim, K., et al., *Bias Field Inconsistency Correction of Motion-Scattered Multislice MRI for Improved 3D Image Reconstruction*. Ieee Transactions on Medical Imaging, 2011. **30**(9).
- [5] Wels, M., et al., *A discriminative model-constrained EM approach to 3D MRI brain tissue classification and intensity non-uniformity correction*. Physics in Medicine and Biology, 2011. **56**: p. 3269.
- [6] Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), 1977: p. 1-38.
- [7] Pham, D.L. and J.L. Prince, *An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities*. Pattern Recognition Letters, 1999. **20**(1): p. 57-68.

- [8] Van Leemput, K., et al., *Automated model-based bias field correction of MR images of the brain*. Medical Imaging, IEEE Transactions on, 1999. **18**(10): p. 885-896.
- [9] Gispert, J.D., et al., *Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error*. Human Brain Mapping, 2004. **22**(2): p. 133-144.
- [10] Wells III, W., et al., *Adaptive segmentation of MRI data*. Medical Imaging, IEEE Transactions on, 1996. **15**(4): p. 429-442.
- [11] Guillemaud, R. and M. Brady, *Estimating the bias field of MR images*. Medical Imaging, IEEE Transactions on, 1997. **16**(3): p. 238-251.
- [12] Manjón, J.V., et al., *A nonparametric MRI inhomogeneity correction method*. Medical Image Analysis, 2007. **11**(4): p. 336-345.
- [13] Ashburner, J. and K.J. Friston, *Voxel-based morphometry--the methods*. Neuroimage, 2000. **11**(6): p. 805-821.
- [14] Lin, M., et al., *A new bias field correction method combining N3 and FCM for improved segmentation of breast density on MRI*. Medical Physics, 2011. **38**: p. 5.
- [15] Gao, G., *A Parzen-Window-Kernel-Based CFAR Algorithm for Ship Detection in SAR Images*. Geoscience and Remote Sensing Letters, IEEE, 2011(99): p. 556-560.
- [16] Csébfalvi, B., *An evaluation of prefiltered B-spline reconstruction for quasi-interpolation on the body-centered cubic lattice*. Visualization and Computer Graphics, IEEE Transactions on, 2010. **16**(3): p. 499-512.
- [17] Lee, J.D., et al., *MR image segmentation using a power transformation approach*. Medical Imaging, IEEE Transactions on, 2009. **28**(6): p. 894-905.
- [18] Hou, Z., *A review on MR image intensity inhomogeneity correction*. International Journal of Biomedical Imaging, 2006. **2006**(49515): p. 1-11.

SESSION

BIOMETRICS + FACE RECOGNITION, EXPRESSION DETECTION, HUMAN DETECTION

Chair(s)

TBA

Finger Vein Recognition in Row and Column Directions Using Two Dimensional Kernel Principal Component Analysis

Sepehr Damavandinejadmonfared, Vijay Varadharajan

Advanced Cyber Security Research Centre
Dept. of Computing, Macquarie University
Sydney, Australia

Abstract - In this paper, a whole identification system is introduced for finger vein recognition. The proposed algorithm first maps the input data into kernel space, then; Two Dimensional Principal Component Analysis is applied to extract the most valuable features from the mapped data. Finally, Euclidian distance classifies the features and the final decision is made. Because of the natural shape of human fingers, the image matrixes are not square, which makes it possible to use kernel mappings in two different ways-along row or column directions. Although, some research has been done on the row and column direction through 2DPCA, our argument is how to map the input data in different directions and get a square matrix out of it to be analyzed by Two Dimensional Principal Component Analysis. In this research, we have explored this area in details and obtained the most significant way of mapping finger vein data which results in consuming the least time and achieving the highest accuracy for finger vein identification system. The authenticity of the results and the relationship between the finger vein data and our contribution are also discussed and explained. Furthermore, extensive experiments were conducted to prove the merit of the proposed system.

Keywords: Biometrics, finger vein recognition, 2-D Principal Component Analysis, Kernel Principal Component Analysis (KPCA).

1 Introduction

Traditionally, Private information is considered as passwords and Personal Identification Numbers (PINs) among the society, which is easy to use but vulnerable to the risk of exposure and being stolen or forgotten. Biometrics[1][2], however, has been attracting researchers' and industry's attention more and more as it is believed that biometrics is a promising alternative to the traditionally used password or PIN based authentication techniques[3]. Nowadays, there are several different biometrics systems under research such as face recognition, finger print, palm print, voice recognition, iris recognition and so on[1][4][5][6][7]. There are two main challenges in terms of biometrics systems. The first one is that

the main element by which the identity is verified or identified is accessible and forgeable, and the second one is that the rate of reliability of the mentioned systems in terms of having a satisfactory accuracy rate is not acceptable. For instance, finger and palm prints are usually frayed; iris images and voice signature are easily forged; face recognition could be considered difficult and unreliable when there are occlusions or face-lifts. Finger vein recognition[8][9][10][11], however, is more secure and convenient and has none of the mentioned drawbacks because of the following three reasons: (1) human veins are mostly invisible and located inside the body; therefore it is difficult to be forged or stolen. (2) It is more acceptable for the user as capturing finger-vein images is non-invasive and contactless. (3) The finger-vein data can only be captured from a live individual. It is thus a convincing proof that the subject whose finger-vein[12] is successfully captured is alive .

Because the data in finger vein recognition is "Image", there are several methods for analyzing and classifying the images in such a recognition system. Principal Component Analysis (PCA)[3][13] is one of the common and powerful methods of pattern recognition and feature extraction which has been used a lot in biometrics. There have been several improvements on PCA such as Kernel PCA (KPCA)[13][14], Kernel Entropy PCA (KECA)[15][16][17] so far. The main drawback of 1-D PCA, however, is that after converting the 2D matrix to 1-D, the dimension of the data is too high which results in having a very time-consuming and even inaccurate system. A highlighted improvement on 1D-PCA is 2D-PCA[18][19][20][21] in which the image matrix is not converted to 1D. This method has two main advantages over 1D-PCA which are being much faster, and having higher accuracy. After proposing 2DPCA, Kernel 2DPCA[20] was introduced in which the data is first mapped to another space using different kernel methods and then 2DPCA is implemented on the mapped data. It is believed that by transforming the data into the appropriate space first and applying 2DPCA on the mapped data, the accuracy rate will have a dramatic increase. As 2DPCA is applied on 2-D image matrixes directly, there has been a great amount of research on the direction of the analysis. 2DPCA can be applied in row direction, column direction, or both. In this paper, however,

different directions of the matrixes for mapping the data into kernel space are argued. We chose Kernel polynomial degree one as mapping function and applied input images in row and column directions. The direction of mapping is important in our system because 2DPCA is applied on the mapped data and extracts as many eigenvectors as the dimension of the mapped data, meaning that if we have a (N by N) matrix, we will have N eigenvectors and their corresponding eigenvalues. For example, assuming our input matrixes are 60*180, applying the mapping function in row and column directions will result in 60*60 and 180*180 matrixes respectively. Applying 2DPCA on 60*60 or 180*180 matrixes will lead to a great difference in accuracies and time duration for the system.

The remaining of this paper is organized as follows:

In Section 2, Image acquisition is explained. In Section 3, Two Dimensional Principal Component is introduced briefly. In section 4, Kernel mapping along row and column direction is explained. In section 5, finger vein recognition algorithm is proposed. In section 6, experimental results on the finger vein database are discussed. Finally, section 7 concludes the paper.

2 Image Acquisition

Based on the scientific fact that the light rays can be absorbed by deoxygenated hemoglobin in the vein [13], absorption coefficient (AC) of the vein is higher than other parts of finger. In order to provide the finger vein images, four low cost prototype devices are needed such as an infrared LED and its control circuit, a camera to capture the images, a micro-computer unit (MCU) to control the LED array, and a computer to process the images [14]. The web-cam consists of an IR blocking filter; hence, it is not sensitive to the infrared (IR) rays. To overcome this issue, an IR pass filter is used to block visible light and pass the infrared light only. Fig. 1 shows 10 samples from a subject. As the figure indicates, the vein pattern is darker than the remaining area in the finger because of the higher absorption of the blood.

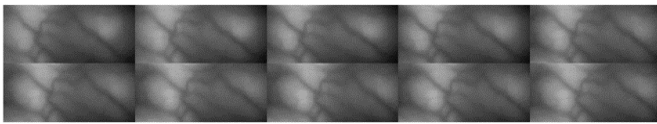


Fig. 1 : a subset of samples captured from a subject

3 Two Dimensional Principal Component Analysis (2DPCA)

The main idea of 2DPCA is to project the image A , which is represented as a random $m \times n$ matrix, onto X that is an n -dimensional unitary column vector:

$$Y = AX \quad (1)$$

Therefore, the projected feature vector of image A is achieved from (1). Finding the appropriate projection vector (X) is the goal. To evaluate the discriminatory power of X , the total scatter of projected samples can be used, which could be characterized by tracing the covariance matrix of projected features vectors. The following criterion is introduced from this point of view:

$$J(X) = tr(S_x) \quad (2)$$

Where S_x represents the covariance matrix of the projected feature vector and $tr(S_x)$ is the trace of S_x . To maximize the criterion in (2), the appropriate projection direction X needs to be found. S_x is introduced by:

$$\begin{aligned} S_x &= E(Y - EY)(Y - EY)^T \\ &= E[(AX) - E(AX)][(AX) - E(AX)]^T \\ &= E[(A - EA)X][(A - EA)X]^T \end{aligned}$$

Therefore:

$$tr(S_x) = X^T [E(A - EA)^T (A - EA)] X \quad (3)$$

Defining the image covariance scatter matrix, we have:

$$G_t = E[(A - EA)^T (A - EA)] \quad (4)$$

We now obtain the $n \times n$ matrix of G_t from all training images, where there are m training images of size $m \times n$. Thus, G_t can be evaluated by:

$$G_t = 1/M \sum_{i=1}^M (A_i - \bar{A}) (A_i - \bar{A})^T \quad (5)$$

Where \bar{A} is the mean matrix of input images and finally we have:

$$J(X) = X^T G_t X \quad (6)$$

First, the $n \times n$ matrix of G_t is calculated from all of the training images. Then, the unitary vectors X are obtained by getting the eigenvector matrix of G_t . This stage decides how many eigenvectors are to be used in the projection of data. To achieve this, the eigenvalues of the corresponding eigenvectors are arranged in a descending order, and a subset of the higher values is selected. Assuming d eigenvectors (with optimal projection axes X_1, X_2, \dots, X_d) are selected, then how to achieve feature extraction and classification stages are explained in the next section.

4 Kernel Mapping Along Row and Column Direction

4.1 Two Dimensional Kernel Principal Component Analysis

The main idea of using kernel function in PCA is that the data is first mapped into another space using a mapping function and then PCA is performed on the nonlinearly mapped data. 2DPCA is better than 1-D PCA in terms of speed and accuracy. The idea of using kernel function in 2DPCA is to improve the accuracy of the system. With N input images, let A_i be i^{th} image, where $i = 1, 2, \dots, N$, and A_i^j be the j^{th} row of the matrix A_i where $j = 1, 2, \dots, n$. The nonlinear mapping is defined as follows:

$$\Phi(A_i) = \begin{bmatrix} \Phi((A_i^1)^T)^T \\ \dots \\ \Phi((A_i^n)^T)^T \end{bmatrix} \quad (7)$$

The total scatter matrix in K2DPCA can be calculated:

$$G_i^\Phi = \sum_{i=1}^N \Phi(A_i) \Phi(A_i)^T \quad (8)$$

Thus:

$$= \sum_{i=1}^n \begin{bmatrix} \Phi((A_i^1)^T)^T \\ \dots \\ \Phi((A_i^n)^T)^T \end{bmatrix} \left[\Phi((A_i^1)^T), \dots, \Phi((A_i^n)^T) \right] \quad (9)$$

$$G_i^\Phi = \sum_{i=1}^N \sum_{j=1}^n \Phi((A_i^j)^T) \Phi((A_i^j)^T)^T$$

In K2DPCA, after achieving G_i^Φ , obtaining projecting axes and the projection and classification procedures are same as in 2DPCA.

4.2 Kernel Mapping in Row and Column Directions & 2DPCA

Equations (7), (8), and (9) demonstrate how the kernel mapping is performed on the input data. Our argument here is that by applying this mapping in different directions (along row and column directions), we will end up having two different data having different dimensions. To further elaborate this, let us assume we have $m \times n$ input matrixes (n is greater than m). By applying the kernel mapping function along row direction (we have m rows meaning that there are m elements for the kernel matrix to be made of) the kernel matrix will be ($m \times m$). In this case, by applying kernel mapping, we reduce the number of input data and its dimension in kernel space which is ($n \times n$). However, if the kernel function is applied along column direction, the kernel matrix will be ($n \times n$). In this case we have expanded the input data to a higher dimensional space

and we have more information to analyse by 2DPCA. Fig. 2 shows the detailed diagram of the kernel mapping and 2DPCA on the mapped data in two different directions. It is observed from the diagram that by applying the kernel function from row or column direction the kernel matrix (K) is squared and with dimension of n or m .

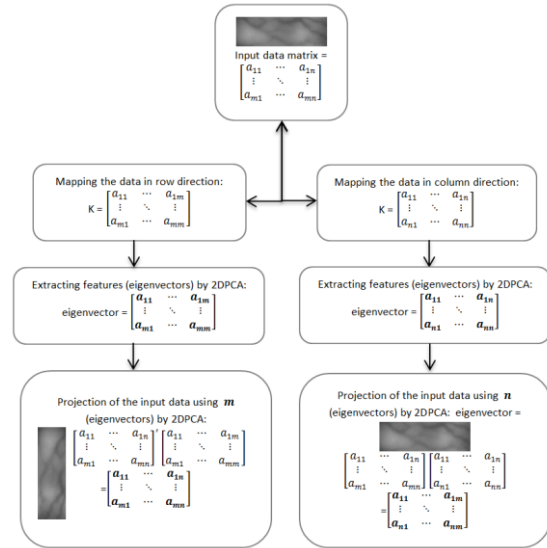


Fig. 2 : Flow diagram of kernel mapping along row and column direction and applying 2DPCA

This argument is indispensable because the dimension of the data affects the output of the 2DPCA greatly. Having higher dimension and more information and features does not guarantee ending up more promising results and higher accuracies. Furthermore, the higher the dimension is the more time-consuming the system is. On the other hand, there has to be a balance between the dimension of the data, the number of used features and the algorithm which is used to analyze the data.

5 Finger Vein Recognition Algorithm

Our proposed finger vein recognition algorithm is explained in this Section. As it is shown in Fig. 3, the algorithm consists of five steps; first step is to extract the region of interest (ROI). Second one is to normalize the images. Third step is to map the data into kernel space along row and column directions which was explained in section 4. In the fourth step, 2DPCA is applied on the data and features are extracted. Last step is to classify the data using Euclidian distance. The flow diagram of the proposed algorithm is indicated in Fig. 3. All steps except for step 3 and 4, which were explained in section 4, are introduced in the following part of this section.

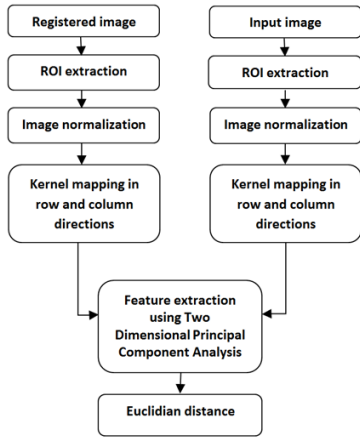


Fig. 3 : Flow diagram of the proposed algorithm

5.1 ROI Extraction

The unwanted black area around the images should be cropped as this area reduces the accuracy and is considered as nothing but noise. To crop images optimally, the used algorithm consists of three major steps. First of all, the edge is detected. Using the detected edges two horizontal lines are determined and the image is cropped horizontally according to the detected lines. Last but not least, the image is cropped vertically at 5% percent from the left border and 15% from the right border.

5.2 Image Normalization

In order to achieve the highest accuracy in least time, images are normalized to smaller size after ROI extraction. It is obvious that the smaller the size is the faster the system is. However, if the size of the image is too small, it may cause too much loss of information as well. Therefore, there has to be a balance between size of the images and the accuracy of the system. Based on our experiments, when using 2DPCA to extract the features, the optimal size of finger vein images resulting in both least time consumption and highest accuracy is 20×60 . Thus, all images are normalized into 20×60 .

5.3 Feature Extraction and Classification Method

As it was mentioned before, Euclidian distance is used as a classifier in this system. Euclidian distance is a very fast method which, we believe, is appropriate for this system because after using kernel map and 2DPCA, the dimension of the data is reduced and therefore Euclidian distance is sufficient to be used.

Given an image sample A , and the optimal projection axes (selected eigenvectors, X_1, X_2, \dots, X_d), the projection will be as follows:

$$Y_k = AX_k, k = 1, 2, \dots, d \quad (10)$$

Using d axes to project the data onto, we will get d projected feature vectors Y_1, Y_2, \dots, Y_d . These vectors are the principal component of the sample image A . Putting these vectors in the form of a matrix, we will get feature matrix of the image A , which is $m \times d$, $B = [Y_1, \dots, Y_d]$.

Then, a nearest neighbor classifier is used to classify the data after transferring all images by 2DPCA and obtaining the feature matrix of them. Considering $B_i = [Y_1^i, Y_2^i, \dots, Y_d^i]$ and $B_j = [Y_1^j, Y_2^j, \dots, Y_d^j]$, the Euclidian distance between them is defined as follows:

$$d(B_i, B_j) = \sum_{k=1}^d \|Y_k^{(i)} - Y_k^{(j)}\|_2 \quad (11)$$

6 Experimental Results on Finger Vein Database

In this section, the experiments conducted on finger vein data are given and explained. Experimental results are explained in two subsections; column direction analysis in experimental setup 1, and row direction analysis in experimental setup 2. Our database consists of 10 samples for each of 100 individuals which results in a total number of 1000 images. In each different part of the experiments, three different types of training and testing were used. 2, 3, and 4 random selected images were used to train each time and respectively, the remaining 8, 7, and 6 images to test. All implementations in each part were repeated as many times as the number of total eigenvectors.

6.1 Experimental Setup-1

To analyze the system in column direction and get the output, we first restate that the images are in dimension of (20×60) meaning that if we map them in column direction, there will be 60 samples with the dimension of 20 to map. The output of such a mapping function will be a matrix with the dimension of (60×60) . By applying 2DPCA on this matrix, there will be 60 eigenvectors extracted with the dimension of 20. In this step, there are 60 different dimensions which could be reduced using projection in 2DPCA meaning that there are 60 different projections using different number of eigenvectors. We conducted the algorithm 60 times in each of three different types (2, 3, and 4 sample to train and 8, 7, and 6 samples to test respectively) of our implementation and calculated the accuracy rate in each point. The obtained results were then gathered and shown in Fig. 4. As it was expected, by adding the number of samples for train, the accuracy goes up no matter which method to use. Another expectation was that by using more eigenvectors the accuracy rate goes higher up to its optimized point, which here is almost near the dimension of 20. It is observed from the Fig. 4 that using 2, 3, and 4 images to train leads to the accuracy rate of around %90, %95, and %97

respectively. Another prime issue is that the time consumed for this experiment is much more than the next as there are 60 eigenvectors in column direction mapping.

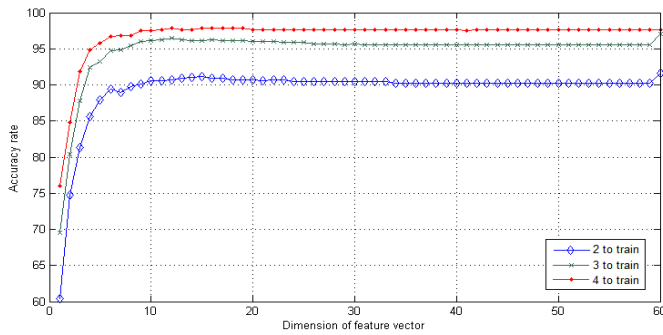


Fig. 4: Accuracy rates obtained using K2DPCA in column direction on finger vein database

6.2 Experimental Setup-2

To analyze the system in row direction the input images were used in a way that each image consists of 20 samples with the dimension of 60 to map. The output of such a mapping function will be a (20×20) matrix. By applying 2DPCA on this matrix, there will be 20 eigenvectors extracted with the dimension of 20. In this step, there are 20 different dimensions which could be reduced using projection in 2DPCA meaning that there are 20 projection manners using different number of eigenvectors. Fig. 5 demonstrate the accuracy rate of the experiments along row direction. Implementing this method using 2, 3, and 4 images to train results to the accuracy rate of around %95, %97, and %99 respectively which is clearly higher than column direction. Not only mapping the input data along row achieves higher accuracy, but also it has less consumption of time as there are only 20 dimensions of data to be reduced.

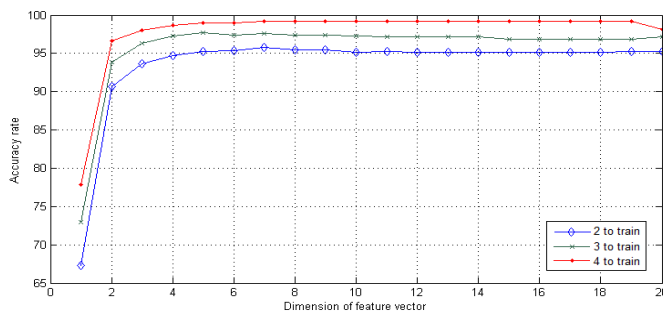


Fig. 5: Accuracy rates obtained using K2DPCA in row direction on finger vein database

In this part, we give a summary of the whole experiments and their corresponding results for the sake of a better comparison. We have chosen the highest accuracies of each method in all implementations and their corresponding dimension of feature vector. All the mentioned information is indicated in Table 1 in addition to the duration of time each algorithm consumed to analyze the data. As the following table shows, the maximum accuracy of the row direction analysis is higher than that of column direction in all the different experiments. Furthermore, it is observed that not only it leads to higher accuracy, but also its dimension of feature vector is much less than that of column method in all implementations implying that the row direction method can be even faster than column direction method in real time system as it reaches the higher accuracy using less feature vectors.

Method	Images to Train	Max Accuracy (%)	Feature Vector No	Duration of Experiment (s)
Column Direction Analysis	2	91.63	60	1324.8372
	3	97	60	1676.553
	4	97.83	15	1805.3797
Row Direction Analysis	2	95.75	7	223.73
	3	97.71	5	311.2961
	4	99.17	7	318.9148

Table 1: Comparison of the proposed algorithm in row and column direction

7 Conclusion

In this paper, we have proposed a new method to enhance the performance of finger vein recognition and analyzed two different aspects of applying it in order to determine the most appropriate one. Our algorithm uses kernel mapping in two different directions to transfer the input data to another space where applying 2DPCA merits the final output of the system. We also used Euclidian distance as classifier in the last step of the algorithm. Extensive experiments were conducted on our database using three different numbers of images for training. Results demonstrate that mapping the data in row direction reaches both having higher accuracy and consuming less time compared to the column direction method meaning that the proposed method has the highest accuracy when mapping the data along the row direction.

8 References

- [1] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Syst.*, vol. 14, no. 1, pp. 4–20, 2004.
- [2] B. Schouten and B. Jacobs, "Biometrics and their use in e-passports," *Image Vis. Comput.*, vol. 27, no. 3, pp. 305–312, Feb. 2009.

- [3] T. S. Beng and B. A. Rosdi, "Finger-vein identification using pattern map and principal component analysis," *2011 IEEE Int. Conf. Signal Image Process. Appl.*, pp. 530–534, Nov. 2011.
- [4] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2228–44, Apr. 2012.
- [5] X. Lu, "Image Analysis for Face Recognition," *East*, pp. 1–37.
- [6] and S. R. Chellapa, C.L. Wilson, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. pp. 705–740, 1995.
- [7] G. Yang, X. Xi, and Y. Yin, "Finger vein recognition based on a personalized best bit map.," *Sensors (Basel)*, vol. 12, no. 2, pp. 1738–57, Jan. 2012.
- [8] W. Song, T. Kim, H. C. Kim, J. H. Choi, H.-J. Kong, and S.-R. Lee, "A finger-vein verification system using mean curvature," *Pattern Recognit. Lett.*, vol. 32, no. 11, pp. 1541–1547, Aug. 2011.
- [9] J. Yang, Y. Shi, and J. Yang, "Personal identification based on finger-vein features," *Comput. Human Behav.*, vol. 27, no. 5, pp. 1565–1570, Sep. 2011.
- [10] J.-D. Wu and S.-H. Ye, "Driver identification using finger-vein patterns with Radon transform and neural network," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5793–5799, Apr. 2009.
- [11] J.-D. Wu and C.-T. Liu, "Finger-vein pattern identification using principal component analysis and the neural network technique," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5423–5427, May 2011.
- [12] P.-Y. Yin, *Pattern Recognition Techniques, Technology and Applications*, no. November 2008. 2008, p. 626.
- [13] K. I. Kim, K. Jung, and H. J. Kim, "Face Recognition using Kernel Principal Component Analysis," *Signal Processing*, vol. 9, no. 2, pp. 40–42, 2002.
- [14] R. M. Ebied, "Feature Extraction using PCA and Kernel-PCA for Face Recognition," *Int. Conf. INFOrmatICS Syst.*, vol. 8, pp. 72–77, 2012.
- [15] P. Hu and A. Yang, "Indefinite Kernel Entropy Component Analysis," *Sci. Technol.*, no. 3, pp. 0–3, 2010.
- [16] R. Jenssen, "Kernel entropy component analysis.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 847–60, May 2010.
- [17] B. H. Shekar, M. Sharmila Kumari, L. M. Mestetskiy, and N. F. Dyshkant, "Face recognition using kernel entropy component analysis," *Neurocomputing*, vol. 74, no. 6, pp. 1053–1057, Feb. 2011.
- [18] P.-C. Hsieh and P.-C. Tung, "A novel hybrid approach based on sub-pattern technique and whitened PCA for face recognition," *Pattern Recognit.*, vol. 42, no. 5, pp. 978–984, May 2009.
- [19] Lin Yang, "A Human Face Recognition Method by Improved Modular 2DPCA," *IT Med. Educ. (ITME), 2011 Int. Symp.*, vol. 2, pp. 7–11, 2011.
- [20] V. D. M. Nhat and S. Lee, "Kernel-based 2DPCA for Face Recognition," *2007 IEEE Int. Symp. Signal Process. Inf. Technol.*, pp. 35–39, Dec. 2007.
- [21] C. Yu, H. Qing, and L. Zhang, "K2DPCA Plus 2DPCA: An Efficient Approach for Appearance Based Object Recognition," *2009 3rd Int. Conf. Bioinforma. Biomed. Eng.*, pp. 1–4, Jun. 2009.

Similarity Measures for Fingerprint Matching

Kareem Kamal A.Ghany¹, Aboul Ella Hassanien² and Gerald Schaefer³

¹Faculty of Computers and Information, Beni Suef University, Egypt

²Faculty of Computers and Information, Cairo University, Egypt

³Department of Computer Science, Loughborough University, U.K.

Abstract—In this paper, we investigate different distance metrics for measuring the similarity between fingerprint templates. In particular, we apply several of them during the matching phase of a fingerprint system, and evaluate the obtained results. Our experiments show the Dice coefficient to give the most convincing results with a matching score of 93%, a false rejection rate of 0.04 and a false acceptance rate of 0.006.

Keywords: Fingerprint matching, similarity measure, distance, Dice coefficient.

1. Introduction

Concepts of similarity and distance are important in many applications. They are for example necessary to measure the similarity of different objects, and thus form an essential part in many pattern recognition applications that involve clustering, classification, recognition, or retrieval. With a large number of similarity measures having been introduced in the literature, selecting an appropriate one for a particular task is crucial, since the success of the related application may depend critically on this choice. Similarity measures vary depending on the data types used [1].

Fingerprint matching refers to finding the similarity between two given fingerprint images. While the choice of matching algorithm depends on which fingerprint representation is being used, the matching algorithm outputs a similarity value that indicates its confidence in the decision that the two images are of the same finger.

In this paper, we compare different similarity measures for fingerprint matching. In particular, in a set of experiments, we evaluate four distance measures – correlation coefficient, Dice coefficient, Sokal and Sneath index, and Simpson index – in the context of the fingerprint recognition approach proposed in [9]. Our results indicate the Dice similarity to give the best results.

2. Similarity Measures

There are many distance metric families used for matching and measuring the similarity between any two objects [2], [3], [4], [5], [7]: Euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance, Sorenson distance, Sorgel distance, Kulczynski distance, Canberra distance, Lorentzian distance, Gower distance, Wave Hedges distance,

Motyka distance, cosine similarity, Jaccard coefficient, Dice coefficient, Hellinger distance, Neyman distance, Divergence distance, Clark distance, correlation coefficient, Mahalanobis distance, etc.

In the following we define the four measure we have used in our experiments.

2.1 Correlation coefficient

The correlation coefficient is defined as [2], [7]

$$\text{Corr}(A, B) = \frac{\sum(a - \mu_a)(b - \mu_b)}{\sqrt{\sum(a - \mu_a)^2 \sum(b - \mu_b)^2}}, \quad (1)$$

and is a measure of statistical dependence between two random variables.

2.2 Dice coefficient

The Dice coefficient is defined as [2], [3], [6], [8]

$$\text{Dice}(A, B) = \frac{2a}{2a + b + c}, \quad (2)$$

where a refers to the features present in both A and B , b the features present only in A and c those present only in B .

2.3 Sokal and Sneath index

The Sokal and Sneath index is defined as [3], [6], [8]

$$\text{SokalSneath}(A, B) = \frac{a}{a + 2b + 2c}. \quad (3)$$

2.4 Simpson index

The Simpson index is defined as [3], [8]

$$\text{Simpson}(A, B) = \frac{a}{\min(a + b), (a + c)}. \quad (4)$$

3. Fingerprint matching system

Our experiments are based on the fingerprint matching system introduced in [9], which we summarise in the following. An overview of the approach is given in Fig. 1.

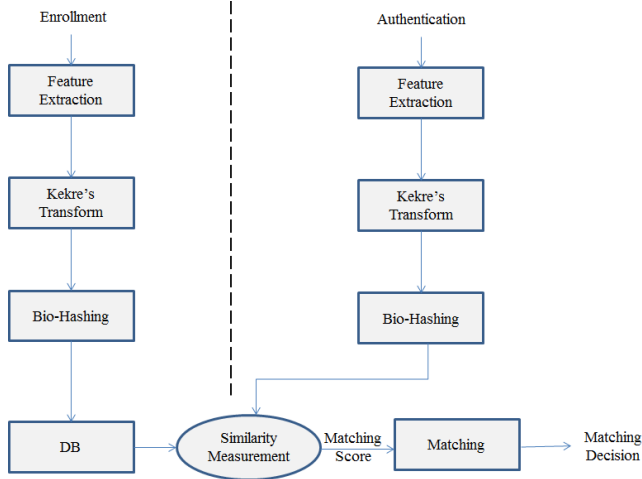


Fig. 1: Fingerprint matching architecture.

3.1 Minutiae extraction

Fingerprint matching typically relies on minutiae extraction. Algorithm 1 details the technical steps of applying the principal curves approach for this purpose.

Algorithm 1 Fingerprint minutiae extraction from principal curves.

for $i = 1 : n$. **do**

Obtain graph $G(vs)$, where V is a set of vertices, $V = \{v_1, v_2, \dots, v_n\}$

$S = \{(v_{i1}, v_{j1}), \dots, (v_{ik}, v_{jk})\} = \{S_{i1j1}, \dots, S_{ikjk}\}$ is a set of edges.

Minimise a penalised distance function $E(G) = \Delta(G) + \lambda P(G)$.

if a single point is found in the extracted edges set **then**
Consider it as an ending of a simple ridge.

else

Consider it as an ending of a ridge bifurcation.

end if

Filter the ridge endings and ridge bifurcations obtained in the extraction step.

end for

3.2 Feature extraction using Kekre's transform

Kekre's transform matrix [9] can be of any size $N \times N$, which need not to be an integer power of 2. We find that all diagonal and upper diagonal elements of this matrix are 1, but that the lower diagonal part except the elements just below diagonal are 0. The general matrix for Kekre's

transform can be given as

$$K_{n \times n} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ -N+1 & 1 & 1 & \dots & 1 & 1 \\ 0 & -N+2 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & -N+(N-1) & 1 \end{bmatrix} \quad (5)$$

with the terms K_{xy} of the matrix calculated as

$$K(x, y) = \begin{cases} 1, & x \leq y \\ -N + (x - 1), & x = y + 1 \\ 0, & x > y + 1 \end{cases} \quad (6)$$

Applying Kekre's transform on an $N \times N$ image, the number of required additions is $2N(N-1)$, and the number of required multiplications is $(N-1)$.

Kekre's transform is applied to the row mean and column mean vectors of the image using Algorithm 2 to obtain Kekre transform row mean and Kekre transform column mean feature vectors respectively. The generated transform coefficients are used as feature vectors of the image. Features for all images in a database can thus be obtained and stored in feature tables for fingerprints images retrieval and matching.

Algorithm 2 Application of Kekre's transform.

for x^t in X , where x is the sample **do**

Calculate $Rv = [\text{Avg}(\text{Row } 1), \text{Avg}(\text{Row } 2), \dots, \text{Avg}(\text{Row } n)]$, where Rv is the row mean vector.

Calculate $Cv = [\text{Avg}(\text{Col } 1), \text{Avg}(\text{Col } 2), \dots, \text{Avg}(\text{Col } n)]$, where Cv is the column mean vector.

Multiply Rv by Kekre's Transform matrix.

Multiply Cv by Kekre's Transform matrix.

Calculate the Euclidean distances from Rv, Cv .

end for

3.3 Bio-hash function

We use a hybrid system that combines a transformation based approach and a biometric cryptosystem approach to increase the performance of protecting the biometrics template process.

Given n minutia points k_1, k_2, \dots, k_n , we can construct the following m symmetric hash functions:

$$\begin{aligned} h_1(k_1, k_2, \dots, k_n) &= k_1 + k_2 + \dots + k_n & (7) \\ h_2(k_1, k_2, \dots, k_n) &= k_1^2 + k_2^2 + \dots + k_n^2 \\ &\dots \\ h_m(k_1, k_2, \dots, k_n) &= k_1^m + k_2^m + \dots + k_n^m \end{aligned}$$

We calculate, in Algorithm 3, the three parameters, translation t

$$t = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}, i = 1, 2, 3, \dots, n, \quad (8)$$

rotation r

$$r = \frac{\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}, i = 1, 2, 3, \dots, n, \quad (9)$$

and error rate E

$$E = \sum_i (y_i - (t + r x_i))^2, i = 1, 2, 3, \dots, n. \quad (10)$$

Algorithm 3 Symmetric hash function using linear least squares

Input: The Euclidean distance between minutiae points with a selected reference point for both the normal and transformed minutiae.

Output: The transformation t , rotation r , and error rate E .

for $i = 1 : n$ **do**

$h'_1 = r h_1 + t$, where (h) is the normal distance, (h') is the transformed distance, and $X_i = \text{sum}(h_i)$, $Y_i = \text{sum}(h'_i)$.

end for

for $i = 1 : n$ **do**

Compute t using Eq. (8).

Compute r using Eq. (9).

Compute E using Eq. (10).

end for

3.4 Matching phase

During authentication, the query is used to recover the original biometric template from the secure sketch and the exact recovery of the original biometric data is verified to authenticate a user. Also, new hash values are produced by the scanner and are matched to those stored in the database. Matching can in principle be performed on both client and server side, and is conducted using hashed features instead of the original template.

Authentication is then depending on the matching score and a pre-specified threshold, which decided whether two fingerprints are from the same finger. The matching score is obtained based on the choice of similarity measure, while the threshold can be set depending on the type of application, e.g. to a more restrictive value for a criminal identification application compared to civilian applications.

4. Experimental Results

We obtained experimental results based on the four chosen similarity measures, i.e. correlation coefficient, Dice coefficient, Sokal and Sneath index, and Simpson index.

We applied these to each of ten fingerprints for ten persons with each fingerprint having five different templates stored in the databases (for a sample see Fig. 2).



Fig. 2: Sample of fingerprint templates used in the experiments.

As performance measure we utilised, the false rejection rate (FRR)

$$FRR = \frac{\# \text{ false rejects}}{\# \text{ genuine matches}}, \quad (11)$$

and the false acceptance rate (FAR)

$$FAR = \frac{\# \text{ false accepts}}{\# \text{ imposter matches}}, \quad (12)$$

as well as the matching score, all calculated for each of the approaches.

As we can see from Table 1, the results can differ quite significantly depending on which similarity measure was chosen. Overall we can notice, that using the Dice coefficient gave the best performance with a matching score of 93% for the same and of 89% for different fingers respectively. The FAR and FRR based on the Dice coefficient were found to be 0.04 and 0.006 respectively, giving a verification rate (defined as $1 - FAR$) of 0.96 and system security (defined as $1 - FAR$) of 0.994 respectively.

Table 1: Matching score results for all similarity measures.

similarity measure	different finger	same finger
Correlation coefficient	0.89	0.92
Sokal and Sneath index	0.18	0.19
Dice Coefficient	0.89	0.93
Simpson index	0.47	0.48

5. Conclusions

In this paper, we have, in the context of a hybrid biometric approach for matching fingerprint templates, evaluated different similarity measures for fingerprint matching. The employed fingerprint matching system first uses a principal curves approach to extract fingerprint features, and then

applies Kekre's transform for fingerprint template transformation. In addition, a bio-hash function for template protecting is utilised. Four different similarity measures have been tested and the experimental results show the Dice coefficient to provide the best performance with a matching score of 93%.

References

- [1] S. Bandyopadhyay and S. Saha, "Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications", Springer, 2013.
- [2] S.D. Bharkad and M. Kokare, "Performance evaluation of distance metrics: Application to fingerprint recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 6, 2011.
- [3] S.S. Choi, S.H. Cha and C.C. Tappert, "A Survey of Binary Similarity and Distance Measures", *Journal on Systemics Cybernetics and Informatics*, vol. 8, no. 1, pp. 43-48, 2010.
- [4] G. Salton and M. McGill, "Introduction to modern information retrieval", McGraw-Hill, 1983.
- [5] L. Leydesdorff, "Similarity Measures, Author Cocitation Analysis, and Information Theory", *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 769-772, 2005.
- [6] B. De Baets, S. Janssens and H. De Meyer, "On the transitivity of a parametric family of cardinality-based similarity measures", *International Journal of Approximate Reasoning*, vol. 50, pp. 104-116, 2009.
- [7] A.A. Goshtasby, "Similarity and Dissimilarity Measures", in "Image Registration: Principles, Tools and Methods", Springer, 2012.
- [8] D. Sanchez and M. Batet, "Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective", *Journal of Biomedical Informatics*, vol. 44, pp. 749-759, 2011.
- [9] K.K.A. Ghany, H.A. Hefny, A.E. Hassanien and M.F.Tolba, "Kekre's Transform for Protecting Fingerprint Template", 13th International Conference on Hybrid Intelligent Systems, pp. 186-191, 2013.

Edge Histogram Descriptor for Finger Vein Recognition

Yu Lu¹, Sook Yoon², Daegy Hwang¹, and Dong Sun Park²

¹Division of Electronic and Information Engineering, Chonbuk National University, Jeonju, South Korea

²Department of Multimedia Engineering, Mokpo National University, Jeonnam, South Korea

Abstract - Edge histogram descriptor (EHD) is an efficient texture representation method originally proposed in MPEG-7 to express the local edge distribution in an image. To efficiently utilize the edge and orientation features of rich veins located inside a finger, in this paper, we propose a finger vein recognition method using edge histogram descriptor. Different from the original usage that divides the image space into 4×4 sub-images, we investigate the relationship between finger vein recognition performance and partition style of input image. The optimal parameter is searched for final recognition. Additionally, the nearest neighbor classifier with Euclidean distance metric is employed for matching. Experimental results on an available finger vein database, MNCBNU_6000, show that the proposed method performs better than those using state-of-the-art algorithms.

Keywords: finger vein recognition, edge histogram descriptor, orientation feature

1 Introduction

Automatic personal identification using biometric characteristics is increasingly developed over the last two decades. However, no biometric has been proved to be perfectly reliable, robust and secure. The defects of traditional biometrics and the growing demand for more friendly and secured biometrics systems have motivated researches to explore new biometric features and traits [1].

Finger vein recognition, being convenient, non-invasive and with high security, has attracted considerable attentions in the past decade. Compared with the traditional biometrics (e.g. fingerprint, facial image, iris, gait, etc), finger vein recognition has the benefits of high anti-counterfeiting, low cost, easy data acquisition with contactless operation, universality and liveness [2, 3]. Furthermore, since the veins are located internally within the living body, finger vein identification system is less affected by the outer skin surroundings (skin disease, humidity, dirtiness, etc). In contrast to the hand vein- or palm vein-based recognition system, finger vein has the advantage of smaller size of imaging device. Hence, finger vein recognition is considered to be one of most promising solutions for personal identification in the future [4].

Human vision is sensitive to edge features for image perception so that edges in an image are considered as an

important feature to represent the content of an image [5]. Histogram, which is invariant to image translation and rotation, is the most commonly used structure to represent the local and global feature composition of an image. Using histogram to represent the edge distribution can describe the frequency and directionality of the brightness changes in an image [6]. An effective edge histogram descriptor (EHD) is proposed in MPEG-7 to express the local edge distribution in the image. Since EHD aims to describe the local edge distribution in efficient storage of metadata, only 80 histogram bins are contained in the edge histogram. The local histogram only using 80 bins may not be sufficient to represent the global features of the edge distribution. To improve the problem, Park et al. [5, 6] proposed an efficient use of local edge histogram descriptor. Semi-global and global edge histograms were generated from the local histogram bins to describe the global edge distribution in an image. Then, the local, semi-global, and global histogram bins were concatenated to represent the edge distribution of an image. The effectiveness using EHD for feature representation has been proved in the applications of image retrieval [5] and face recognition [7].

The orientation features and edges are abundant in finger vein images, since the veins are located and developed within the finger with random orientation [8]. Different finger vein image brings different blood vessel network, which produces difference of their edge histograms. Hence, to efficiently utilize the edge and orientation features, in this paper, we present a finger vein recognition method using edge histogram descriptor. The original idea in edge histogram operator [5, 6] is to divide an image into nonoverlapping sub-images. We believe the best performance may be achieved with other partition style. Here, the performance obtained by dividing the finger vein image into several numbers of sub-images is evaluated to search the optimal way of image partition. Compared with the existing methods using such orientation features as local binary pattern code (LBPC) [9], local direction code (LDC) [10], Gabor filter [11], and Steerable filter [12], the proposed method shows its superiority with lowest equal error rate (EER) value.

The rest of this paper is structured as follows. Section 2 briefly introduces the preprocessing for ROI localization. Finger vein feature representation using local edge histogram descriptor is reported in Section 3. Section 4 provides

experimental results that verify the proposed algorithm. Finally, conclusion is given in Section 5.

2 ROI localization

Image preprocessing is a crucial process in a finger vein identification system. This process always includes ROI localization, image denoising, and image enhancement. Since the acquired images in our established database have good image quality, image denoising and image enhancement are not necessary in our finger vein identification system. Hence, image preprocessing in this paper only contains ROI localization. To correctly localize the ROI from the acquired image, a robust finger vein ROI localization method based on flexible segmentation is employed in this paper [13].

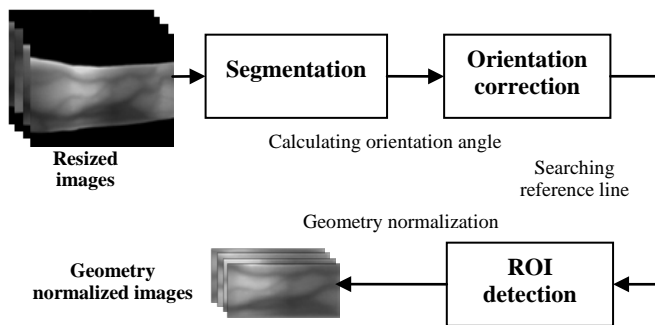


Fig. 1 Block diagram illustration of ROI localization.

Fig. 1 shows the block diagram illustration of the ROI localization method used in this paper, which contains a set of steps, namely segmentation, orientation correction, and ROI detection. In order to reduce the processing time, the acquired image is resized to 120×160 with 'bicubic' interpolation primarily. To get the pure foreground, finger region is segmented from the resized image using an extended edge operator. For the images in the abnormal case, elaborate binarization is utilized to remove the false background caused by the influences from uneven illumination, scattering, and improperly collection. With the middle points obtained from the finger region, the orientation angle can be calculated using least-squares estimation. Afterwards, the resized image is orientation corrected according to the estimated angle. For ROI detection, we extend the method used in [4], since the ROI of the images in our database is defined as a fixed region, based on searching the reference line in the second knuckle. Owing to the width of ROI varies with different finger, geometry normalization is necessary to eliminate the geometry variations. In this paper, all localized ROIs are normalized to 60×128 pixels with 'bicubic' interpolation.

3 Feature extraction and matching

In this paper, EHD is employed as feature representation method for finger vein identification.

3.1 Local edge histogram descriptor

The EHD basically represents the edge distribution using 5 types of edges in each local area, called a sub-image. Usually, an image is divided into 4×4 nonoverlapping blocks, which is shown in Fig. 2. Thus, the image partition always creates 16 equal-sized sub-images regardless of the size of the original image [6]. Edges are grouped into five categories (Fig. 3): horizontal, vertical, 45 diagonal, 135 diagonal and non-directional edges. Thus, for each sub-image, five bins of edge histogram can be obtained, corresponding to the above five categories.

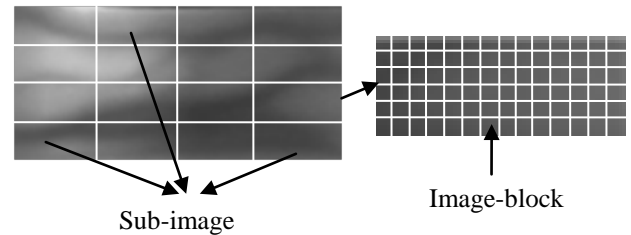


Fig. 2 Definition of sub-image and image-block.

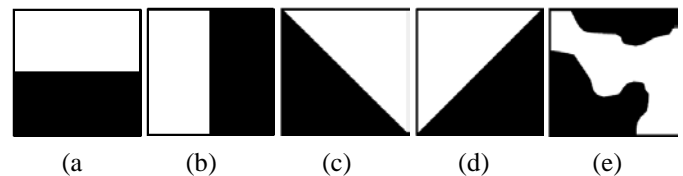


Fig. 3 Five edge types: (a) horizontal edge, (b) vertical edge, (c) 45 diagonal edge, (d) 135 diagonal edge, and (e) non-directional edge.

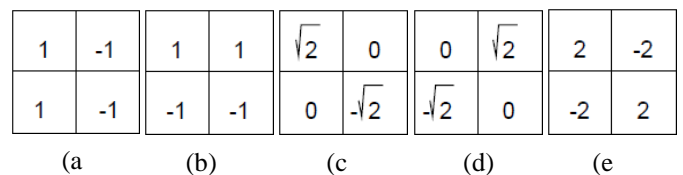


Fig. 4 Five edge filters: (a) horizontal filter, (b) vertical filter, (c) 45 diagonal filter, (d) 135 diagonal filter, and (e) non-directional filter.

To characterize the edge distribution, the sub-image is further divided into small square blocks called image-block (Fig. 2). The number of image-blocks is constant and independent of the original image dimensions. The size of image-block is proportional to the size of original image to deal with the images with different resolutions. Mean values of the four sub-blocks are obtained first. Then they are convolved with five filters shown in Fig. 4. Hence, five directional edge magnitudes are calculated, corresponding to five edge types. If the maximum magnitude value is larger than a threshold, the image-block is considered having the corresponding edge type. After the edge extraction from all the image-blocks, we can obtain the edge distribution of this

sub-image. Thus, we compute five histogram bins for each sub-image. Afterward, each histogram is normalized by dividing each bin with the total number of image-blocks in the sub-image. Since an image is usually divided into 4×4 sub-images, we have total 80 ($4 \times 4 \times 5$) bins for the edge orientation histogram.

3.2 Global and semi-global edge histogram descriptor

Only 80 bins in edge histogram are not sufficient to represent the global edge features. To represent the image with global edge distribution, global and semi-global edge histograms are directly extracted from the local edge histograms in [5, 6]. Since there are 5 edge types, the global edge histogram can be obtained by adding the 16 local edge histograms. Hence, the global edge histogram has five bins. For the semi-global histograms, four connected sub-images (as shown in Fig. 5) are connected to compose a semi-global histogram. For 4×4 sub-images, we can get 13 different clusters. Consequently, we have total 150 (80 (local) + 5 (global) + 65 (semi-global)) histogram bins for each image. Clusters from 1 to 4 and 5 to 8 emphasize the vertical and horizontal edge connectivity, respectively.

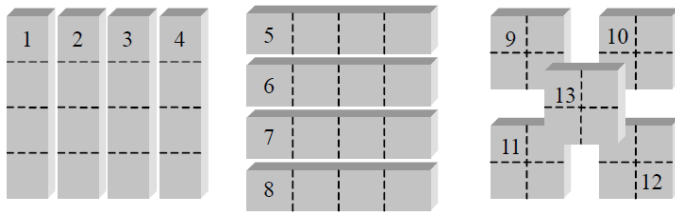


Fig. 5 Clusters of sub-images for semi-global histograms [5].

3.3 Feature representation using EHD

The usage of EHD in [5, 6] just divided an image into 4×4 nonoverlapping blocks. However, we believe a better performance can be obtained if the features contain more edge information. Hence, in this paper, we test the performance using different kinds of partition styles. Furthermore, to get more abundant edge information in each sub-image, each image pixel is considered as a basic unit (image-block) here.

Suppose the finger vein image is divided into $m \times n$ nonoverlapping sub-images. For the local histogram, it has $m \times n \times 5$ bins. For the global histogram, it has 5 bins, no matter how many sub-images there are. For the semi-global histograms, we have $m+n+5$ clusters. Thus, $m+n$ histograms are generated similar with the clusters from 1 to 8 in Fig. 5. It also has 25 bins for 5 histograms as shown from 9 to 13 in Fig. 5. These 5 histograms are generated from 4 sub-images located in each corner (4 corners) and another 4 sub-images located in the middle of the image. Therefore, there are total $m \times n \times 5 + 5 + [(m+n) \times 5 + 5 \times 5]$ bins for each finger

vein image when the input image is divided into more than 4×4 sub-images.

3.4 Matching

There are various similarity metrics to evaluate the similarity from two histograms, such as histogram intersection, log-likelihood ration and chi-square. As the edge orientation histogram is normalized by dividing the total number of image-blocks in each sub-image, in this paper, we employed Euclidean distance for measuring the similarity of two histograms from two images.

4 Experimental results

All the experiments in this part are implemented on our established database which is named MNCBNU_6000 [14, 15]. For equitable comparison, all experiments are performed on the localized ROI, without any post-processing like image denosing and enhancement.

4.1 Finger vein dataset

MNCBNU_6000 consists of finger vein images captured from 100 volunteers, who are students and professors in CBNU from Asia, Europe, Africa, and America, coming from 20 different countries. The ages of volunteers are from 16 to 72 years old. Statistical information of the nationality, age, gender, and blood type of each volunteer is available for deep analysis on the finger vein image. Since the length of the thumb and the little finger is too short, compared with other three fingers, each subject was asked in the capturing process to provide images from his or her index finger, middle finger, and ring finger of both hands in a standard office environment (rather than a darkroom). The collection for each of the 6 fingers is repeated 10 times to obtain 10 finger vein images. For each image collection, the subject was asked to input his or her finger optionally. Our finger vein database is composed of 6,000 images. Each image is stored in "bmp" format at 480×640 pixels size.

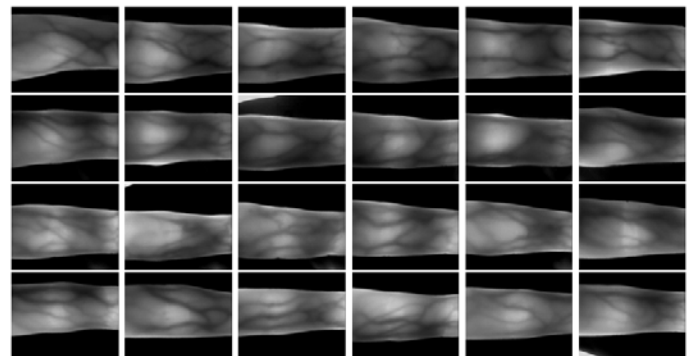


Fig. 6 Some finger vein image samples in MNCBNU_6000. Each row represents six images from six captured fingers of a person.

4.2 Searching optimal parameters

There are two parameters that will affect the performance of finger vein system using the proposed method. One is the number of sub-images. The other one is the threshold that is used for comparing with the maximum magnitude value. To investigate the optimal parameters, we design two experiments with performance evaluation using EER, which is the value where the False Accept Rate (FAR) is equal to the False Reject Rate (FRR). In addition, the receiver operating characteristic (ROC) curve generated by adjusting the matching threshold is also created for comparison. In both of these two experiments, five finger vein images from one individual are selected as the training set, while the rest five images are used as the test set. Hence, the training database and testing database are both composed of 3,000 images. Each finger is considered as an individual.

The first experiment is designed for searching the optimal threshold, while the partition style is 4×4 in this experiment. Fig. 7 shows the EER values with varying threshold. As mentioned above, the threshold is applied for deleting some useless points with weak edge. Thus, the matching performance enhances a little when the threshold increases from 0 to 2. However, with further increasing the threshold, some useful edge information is lost, which results in the enhancing EER values. Hence, the threshold is fixed to be 2 in the later experiments.

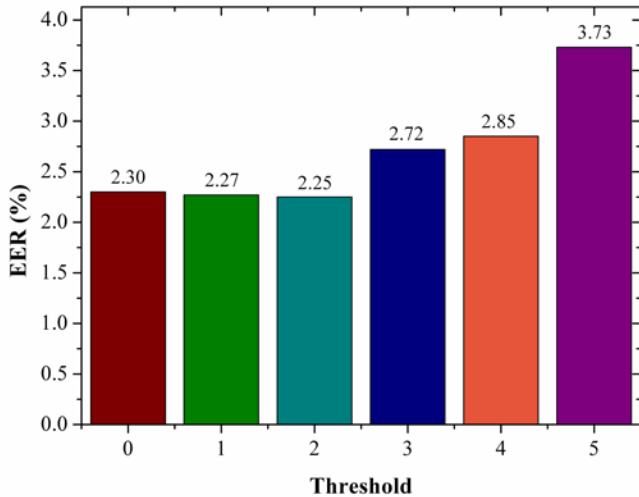


Fig. 7 EERs with increasing threshold.

The second experiment aims to investigate the optimal partition type. Table 1 shows the EERs with the varying partition styles. With the increasing number of sub-images, the EER values are decreasing with different extents. It shows the lowest EER of 1.34% when an image is divided into 8×8 sub-images. However, with the further increasing of the number of partitioned sub-images, the EER value enhances a little. The EER value is 1.46% when the input image is divided into 9×9 sub-images, which is larger than that of

dividing image into 8×8 sub-images. Additionally, the larger number of sub-images would cause longer processing time for feature extraction and matching. Hence, the optimal partition style is dividing a ROI image into 8×8 nonoverlapping sub-images.

TABLE I. EER VALUES WITH VARYING PARTITION STYLES

$m \times n$	Number of bins	EER
4×4	150	2.25%
4×6	200	2.08%
5×5	205	1.78%
6×4	200	1.64%
6×6	270	1.57%
6×8	340	1.44%
8×6	340	1.49%
8×8	430	1.36%
9×9	525	1.46%

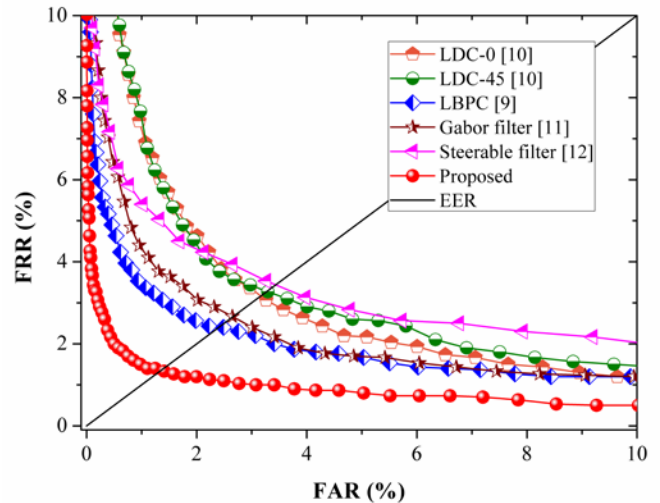


Fig. 8 ROC curves obtained using different methods.

4.3 Matching performance comparison with existing methods

In order to ascertain the performance improvement using the proposed method, state-of-the-art algorithms such

as LBPC [9], LDC [10], Gabor filter [11], and Steerable filter [12] are implemented for comparison.

Fig. 8 shows the ROC curves using different feature representation methods. All these methods extract the directional features for image representation. It is clearly illustrated that the proposed method outperforms the existing methods such as LLBP [9], LDC [10], Gabor filter [11], and Steerable filter [12]. This attributes the efficient extraction orientation features by using EHD. Furthermore, edge histograms extracted from 8×8 sub-images can further effectively represent the orientation and edge features in a finger vein image, other than that extracted from 4×4 sub-images.

5 Conclusions

In this paper, we proposed a finger vein recognition method using edge histogram operator. Instead of dividing the input image into 4×4 sub-images, we investigated the performance with different non-overlapping partition styles. Experimental results, performed on our established finger vein database MNCBNU_6000, demonstrate that the optimal partition style is dividing the ROI image into 8×8 non-overlapping sub-images. In addition, the proposed method performs better than state-of-the-art algorithms that use directional features. In the future, we will devote ourselves to the investigation of new edge operators with 8 orientations and apply that in EHD.

6 Acknowledgment

This work is supported by the State Scholarship Fund organized by the China Scholarship Council (CSC). This work is also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2013778). This work was also supported by the Brain Korea 21 PLUS Project, National Research Foundation of Korea.

7 References

- [1] A. Kumar, and Y.B. Zhou, "Human identification using finger images," *IEEE Transactions on Image Processing*, vol.21, pp:2228-2244, 2012.
- [2] J. C. Hashimoto, "Finger vein authentication technology and its future," *Symposium on VLSI Circuits Digest of Technical Papers*, pp. 5-8, Honolulu, US, 2006.
- [3] T. Yanagawa, S. Aoki, and T.Ohyama, "Human finger vein images are diverse and its patterns are useful for personal identification," *Kyushu University MHF Preprint Series*, Kyushu, Japan, pp. 1-7, 2007.
- [4] J. F. Yang, and Y. H. Shi, "Finger-vein ROI localization and vein ridge enhancement," *Pattern Recognition Letters*, vol. 33, pp. 1569-1579, 2012.
- [5] D. K. Park, Y. S. Jeon, C. S. Won, and S. J. Park, "Efficient use of local edge histogram descriptor," *In Proceedings of the 2000 ACM workshops on Multimedia*, New York, USA, pp. 51-54, 2000.

- [6] C. S. Won, D. K. Park, and S. J. Park, "Efficient use of MPEG-7 edge histogram descriptor," *ETRI Journal*, vol. 24, pp. 23-30, 2002.
- [7] S. Rahman, S. M. Naim, A. A. Farooq, and M. M. Islam, "Performance of PCA based semi-supervised learning in face recognition using MPEG-7 edge histogram descriptor," *Journal of Multimedia*, vol. 6, pp: 404-415, 2011.
- [8] Y. Lu, S. J. Xie, S. Yoon, and D. S. Park, "Finger vein identification using polydirectional local line binary pattern," *In Proceeding of International Conference of ICT Convergence*, Jeju, Korea, pp. 61-65, 2013.
- [9] E. C. Lee, H. C. Lee, and K. R. Park, "Finger vein recognition using minutia-based alignment and local binray pattern-based feture extraction," *International Journal of Imaging Systems and Technology*, vol. 19, pp. 179-186, 2009.
- [10] X. J. Meng, G. P. Yang, Y. L. Yin, R. Y. Xiao, "Finger vein recognitoin based on local directional code," *Sensors*, vol. 12, pp. 14937-14952, 2012.
- [11] S. J. Xie, J. C. Yang, S. Yoon, Y. Lu, and D. S. Park, "Guided Gabor filter for finger vein pattern extraction," *In Proceeding of Eighth International Conference on Signal Image Technology and Internet Based Systems*, Naples, Italy, pp. 118-123, 2012.
- [12] J. F. Yang, and X. Li, "Efficient finger vein localization and recognition," *In Proceedings of 21th International Conference on Pattern Recognition*, Tsukuba, Japan, pp. 1148-1151, 2010.
- [13] Y. Lu, S. J. Xie, S. Yoon, J. C. Yang, and D. S. Park, "Robust finger vein localization based on flexible segmentation," *Sensors*, vol. 13, pp. 14339-14366.
- [14] Y. Lu, S. J. Xie, S. Yoon, Z. H. Wng, and D. S. Park, "An available database for the research of finger vein recognition," *In Proceedings of 6th International Congress on Image and Signal Processing*, Hangzhou, China, pp. 386-392, 2013.
- [15] MNCBNU_6000 database:
<http://multilab.chonbuk.ac.kr/>

Evaluation of the Impact of Noise on Iris Recognition Biometric Authentication Systems

Abdulrahman Alqahtani

Department of Computer Sciences, Florida Institute of Technology
Melbourne, Florida, 32901

Email: aalqahtani2011@my.fit.edu

Abstract—Iris recognition is a particular type of biometric system that uses pattern-recognition techniques on images of irides to uniquely identify an individual. Here I provide evidence that the iris has a unique pattern which gives us reliable form of identification. In this paper, we will discuss the image quality of irises during iris recognition processing. We used the CASIA-IrisV3-Interval of iris database as input in our system. In this paper, we focused on studying and implementing image quality for each image of CASIA-IrisV3-Interval database by adding noise to that database. We used the Hamming Distant technique (HD) to compare each image on CASIA Database with other images in the same database before and after adding noises. When we get the results Of Hamming distance (HD), we will calculate the TPR and FPR to build ROC Carves for evaluating the image quality of CASIA images.

Keywords: iris recognition, Biometric authentication, Iris recognition, image quality, Evaluation of the impact of noise on iris recognition

I. INTRODUCTION

The use of Iris Biometric is one of the important and successful traits used for the authentication and identification of people for different legal or other critical issues. Initially, this technique was used to recognize an Afghan girl falsely after 18 years with the help of her iris patterns. After that the technique was used widely for authentication and identification for different purposes [1]. There are no two irises which are the same in their mathematical detail—even between one's own left and right eye, or between identical twins and triplets [2]. Biometrics systems have come into existence with the revolution of information technology and computer systems. The main purpose of biometrics is to identify individuals and biometrics recognizes the identity of an individual rather than what the person has [3]. Biometrics is not only a system to recognize the identity of an individual, rather it can also be used in our society to reduce fraud, for convenience, and to make our society safer [4]. Biometrics is usually evaluated on the basis of the biometrics recognition performance. The quality and ruggedness of the sensors are two factors that effect the biometrics recognition performance [5]. There are basically two errors in the biometrics system i.e.

false rejection rate FRR (rejection of the client) and the false acceptance rate FAR (acceptance of an impostor) [5]. There are number of developments that are responsible for the growth of the biometrics industry. e.g. the development of software standards, BioAPI, Biometric Service Providers and Biometric Information Records [6].

There are many challenges for the Iris Recognition System (IRS). Our experiment focuses on studying and implementing iris recognition, comparing each image with others in same database (to get the image quality for each image) by using the Chinese Academy of Sciences - Institute of Automation (CASIA) database for testing our project [7].

In this project, we have a number of sub-systems corresponding to each stage of iris recognition authentication systems. These stages are as follows: Segmentation, Normalization, Feature encoding and Matching. The Results of this IRS are

- Iris segmentation templates.
- Iris normalization templates.
- Polar Mask templates.
- Polar Noise templates.

We used the Hamming Distance (HD) technique to match between the iris images of the CASIA database. The IRS has a great advantage. When we compare the IRS to other visual recognition systems there is huge variability. The Iris pattern between individuals varies to such a degree, that it allows large databases to search with reduced number of false matches [7].

According to recognition of human Iris patterns for biometric identification that was presented by L. Masek, the first step of iris recognition is to isolate practical iris regions in a digital eye image. The imaging qualities of eye imaging will predict the quality of segmentation success. After that, the normalization process produces iris regions which contain the same constant dimension, such that two photographs of the same iris, under different conditions, will have characteristic features at the same spatial locations. In feature encoding and matching, the significant features of the iris must be encoded so that comparisons between

templates can be determined. Most iris recognition systems make use of a band pass decomposition of the iris image to create a biometric template [8].

II. BACKGROUND

A. Biometric authentication

Biometrics is basically a terminology that has been derived from the two Greek words, bio, which means life, and metrics which means to measure. It is defined as a method to recognize individuals on the basis of their psychological and behavioral methods [3].

This method of biometrics is preferred to the traditional method of passwords and pin numbers because of its accuracy and authenticity [9]. The word biometrics denotes automatic identification of persons grounded on their interactive and organic features. Numerous biological, along with social, biometric individualities have been used. For example, Fingerprint, Palm Print, Face, Iris, Retina, Ear, Voice, Signature, Gait, Hand, Vein, Odor, DNA, etc., are contingent on various kinds of requests.

Biometric characters are developed relating satisfactory devices and unique topographies are mined to customize a biometric pattern in registration development. Throughout verification, which is also known as authentication procedure or identification, it can be an identification that is touched as an order of verifications and selections. The organization procedures are the additional biometric dimension which is likened beside the kept pattern(s) producing rejection or acceptance [10].

Historically, humans used faces to recognize each other, but with the increase in populations and as the mode of transport increased, the need for recognition increased, which led to emergence of the field of biometrics. The biometric system can either be a verification system, or an identification system [9]. A verification system is basically to confirm a person's claimed identity, while the identity of a person is established in the identification mode. [9].

Applications of biometrics include mobile phones, secure electronic banking, and computer systems security, secure access to buildings, health, credit cards and social services [3]. Commonly used biometrics includes: Infrared thermogram (hand, hand vein, and facial), Gait which is the peculiar way one walks, Keystroke, odor, Ear, Hand print, Retina, Iris, Palm print, Voice, Face, Signature, and DNA [3]. There are basically two phases in the biometrics system; i.e. the recognition phase, and the learning phase [5].

An item under consideration is recorded with the help of sensors when the digital data is available. The data is not used directly; rather, some of the data characteristics are extracted from the digital data first to form template [5]. The responsibility of the learning phase is to create a model, e.g. a statistical model. The recognition phase deals

with the decision to be taken [5]. The three main functions that biometrics can perform is positive identification of the individual, i.e. Positive Identification (does the system know this person?). It is used for large scale identification, i.e. (whether the person is in the data base or not). Lastly, biometrics perform a function of screening, such that it asks the question, Whether this is a wanted person or not? [4]. Biometrics offer huge amount of security and privacy as compared to other methods of identifying individuals. In some cases the biometrics can be considered as a replacement of the technology [5]. There are basically three main applications of biometrics, i.e commercial, government, and forensic applications. The commercial application includes electronic data security, ATMs, physical access control, Internet access, computer network logins, e-commerce and credit cards, etc.

The government applications include correctional facilities, welfare-disbursement, social security, national ID cards, driver's licenses, and passport control, etc. Finally, forensic applications include criminal investigation, paternity determination, terrorist identification, corpse identification and missing children.

The increased threats of terrorism have seen an increased use of biometrics today. A human hand has a combination of different features and they vary significantly from one person to another. Geometric measurement are mostly used as a means of recognitions in commercial systems. Reference pegs are mostly used in geometric measurements for capturing the image of the hand. The most important factor in geometric measurement is a user's acceptability [11].

There is zero risk of the biometrics being lost or forgotten, so the potential threat of intruders is also minimized by the use of the biometrics. Identification is more difficult than verification. Large numbers of comparisons are implementing biometric systems in order to identify individuals. The individual who wishes to remain anonymous can be deprived of their privacy by these biometric systems [5]. Multi-modal biometric systems use the data provided by multiple biometric sources to identify an individual [12]. Information collected from different sources is amalgamated from three levels; i.e., match score level, decision level, and the feature extraction level [12].

B. Iris Recognition

In this paper, we have a number of sub-systems that correspond to each stage of the iris recognition. These stages are: -

- Image acquisition(CASIA-Iris (Chinese academy of sciences-institute of automation) database)
- Segmentation
- Normalization
- Feature encoding

According to the Human Iris Recognition Patterns for Biometric Identification that are presented by L. Masek. Masek's method uses the automatic segmentation system which is established on the Hough transformation method. It can get the location of the circular iris and pupil regions. In the Normalization stage, the extracted iris region (the results of segmentation) is normalized into a rectangular block with constant dimensions to account for imaging inconsistencies. Lastly, to encode the unique pattern of the iris into a bit-wise biometric template, the phase data from 1D Log-Gabor filters must be extracted and quantized into four levels.

The Hamming distance is used for the rating of the iris templates. The couple templates are found to match if the testing of statistical independence has been unsuccessful.

In our project, the inputs to the system are the eye images, and the outputs are the iris templates [8], [13], [14], [15], [16]. Image Acquisition is taking an image from an iris in the initial stage of an iris-based recognition system [3]. In our project, we do not use this technique because we implemented it with the Chinese academy of sciences-institute of automation (CASIA) database.

In fact, an image acquisition captures more than just the iris; e.g. pupil, eyelid, Eyelashes and Sclera (white part of the eye). Also, segmentation isolates the iris from the rest of the eye. There are several techniques and algorithms available

- Hough transform which was implemented by Wildes et al, which we used in our experiment.
- Daugman's algorithm (integro-differential operator) [14] [15] It was submitted in 1993. It was the first method efficiently employed on the biometric system. The data-set of this technique is known integro-differential operator [15].
- Shrinking and expanding active contour methods [17] These methods are unified when localizing inner and outer iris boundaries. First, the pupil region is assessed based on histogram threshold and morphological operations. Afterward, it uses this data to locate the inner iris boundary. Finally, the inner iris boundary is taken as an initial contour to obtain the outer iris boundary.

There are several techniques available for Normalization

- Daugman's rubber sheet model, we used this technique in our project.
- Image registration methods. Image registration is processing of more than one image which is used by the same scene taken at different times, from different sides, or by using different sensors [18] [8].

- Virtual circles are used to ensure accurate location. [19]

The segmented iris region is normalized to eliminate dimensional inconsistencies between iris regions. This is done by implementing a version of Daugman's rubber sheet model [15] [18].

There are several feature extraction and encoding methods and techniques available

- Log Gabor filters, which we used in our experiment. [13]
- Zero crossings of 1D wavelet; a wavelet is a function that is used to build a representation [19].
- Laplacian of Gaussian, it is used for evaluation of the qualities of the iris images [8] [20] [20].

By using 1D Gabor wavelets with the normalized iris pattern, we are able to prove that Feature encoding is based on the polar coordinate on the 2D normalized iris image. [13].

There are several methods and techniques available for matching

- Hamming distance which was implemented by Daugman, which we used in our experiment. [8], [14]
- Weighted Euclidean distance that is employed by Zhu et al. It implemented Weighted Euclidian Distance (WED) (2000) by Zhu et al. It is used to compare the distance between the two templates, particularly if the templates are composed of integer values [21].
- Normalized cross-correlation that is employed by Wildes et al, [20] [22]. It is used for matching parts of the images in many applications. Matching methods based on normalized cross-correlation can handle the scale changes between the two images, where there is translation or rotation [22].

C. Image quality

There are numerous research projects and papers which discussed the image quality of irises, most of them focused on the noise effects and the different algorithms of iris segmentation.

In this paper, we focused on the image quality of iris recognition by comparing each image with other images from the same database, which is CASIA V3 database. As

we mentioned in previous sections, that the image of the eye has to follow all iris recognition stages to obtain the final decision of the iris recognition system (matching or non-matching).

Our aim by using an iris recognition system is to know the rate of image quality for each image by comparing it with others. Then, we build our ROC curve according to Hamming distance (HD) techniques.

Previous research on iris image quality can be divided into two classes: local and global analysis. Zhu uses the employing discrete wavelet decomposition to measure iris quality by analyzing the coefficients of iris texture [23].

Chan et al have classified the iris quality by evaluating the vitality of concentric iris bands acquired from 2-D wavelets [24].

Ma et al. characterized defocus, motion, and occlusion by analyzing the Fourier spectra of local iris regions [25].

Zhang focused on the sharpness of the area between the pupil and the iris [26].

Daugman and Kang described quality by quantification of the energy of high frequencies over all image regions [27], [13].

Most of the previous research evaluation of iris quality involves some traditional segmentation methods [28]. There are four popular image segmentation methods:

- Daugman's method Daugman supposed that both pupil and iris circular form applies an integrodifferential operator.
- Wildes's method This method executes iris contour appropriate in two steps [29] First step converted the image information into a binary edge map. Second step for particular contour parameter values uses the edge points vote.
- Mask's method: Using the Kovese's edge detector, variation is known as Canny edge detector. Then, the next step is applying the circular Hough transform to determine the iris/sclera, and each image correspondent to the iris/pupil.
- Liam and Chekima method: This technique is based on the fact that the pupil is darker than the iris, but the iris is darker than the sclera.

In most cases, the iris images have been taken in less attitudinized imaging conditions, including noise which is localized in some of the iris subparts. Usually, reflections are in the left /right iris and the obstructions are in the upper or lower part of the iris. In the lower and middle-lower signal frequencies, the common feature extraction methods focus on ways to make it more likely that noisy data is included to create the biometric signature.

We should divide the iris into different regions in order to detect the regions that are noise-free, then use it to compare with enrolled regions for accurate recognition.

Hugo Proenca and A. Alexandre, in their experiment cap-

tured Iris images under simulated noncooperative conditions to reduce false rejections. Moreover, they have a free database (UBIRIS), which has distinguishing characteristics of the two free databases (CASIA and UPOL), and the exiting of this data is noise-free. Hugo Proenca and A. Alexandre's experiment has compared the aforementioned four image segmentation methods by using UBIRIS and CASIA [29].

Mayank Vatsa and his group in their experiment proposed that iris indexing algorithms use local and global features to decrease the identification time without compromising accuracy. This algorithm significantly minimizes the computational time without any effect on the accuracy of identification.

Mayank Vatsa and his group discussed the challenge of improving performance by comparing the verification and identification algorithms using these iris databases: CASIA Version 3, ICE2005 and UBIRIS [30].

Peihua and Hongwei's experiment discussed the iris recognition problem with errant capture in non-idealistic imaging conditions. In those cases, the iris recognition is challenged by noisy factors; e.g. the off-axis imaging, pose variation, image blurring, illumination change and occlusion, specular highlights, and noise.

They provided a robust algorithm based on the Random Sample Consensus (RANSAC) to localized non-circular boundaries. Peihua and Hongwei asserted that these methods can be more accurate than Hough transform methods for localization of iris boundaries by selecting the method based on LucaseCkanade algorithm. According to their experiment, one image could divide into small sub-images and this fixes registration problems for every small sub-image by operating the filtered iris image.

Peihua and Hongwei presented their selection method for getting a sub-optimal subset of filters from a family of Gabor filters using UBIRIS.v2 databases. The recognition performance will be greatly improved with a small number of section filters [31].

III. METHODOLOGY

In our experiment , we used the iris recognition system with the CASIA-IrisV3-Interval database then we generated the ROC curve for original database. We added different noises to images of original database for creating different data images with different noises. The ROC curves were generated for each data with its noise. Finally, we compared between ROC curves with and without noises. Overview of our proposed method is shown in Figure 1.

A. Data

We used CASIA V3-Interval. The Chinese academy of sciences-institute of automation (CASIA) database. There

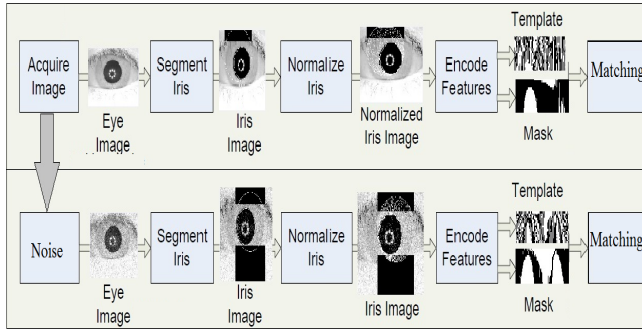


Figure 1: Overview of our proposed method.

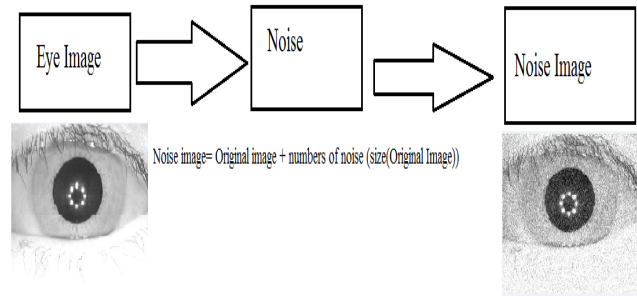


Figure 2: The process of adding noise.

are a few iris databases which have sample images that are available free of charge to the public. These iris databases share a vast amount of iris images, made in different places. Most of these iris databases, when they are constructed, will strive to have a vast collection of quality iris images. [32]. The Center For Biometric and Security Research in China provided this database to the public in order to improve iris recognition system research around the world [7].

The CASIA-IrisV3 contains three types which are known as CASIA-IrisV3-Interval, CASIA-IrisV3-Lamp and CASIA-IrisV3-Twins. CASIA-IrisV3 includes a total of 22,035 iris images that were taken from more than 700 subjects.

All iris images are collected under near infrared illumination, in which all images are 8 bit gray-level JPEG files [7]. In our experiment, we used the CASIA-IrisV3-Interval (Chinese academy of sciences-institute of automation) database. The CASIA-IrisV3- Interval includes a total of 2,639 iris images that were taken from 249 subjects.

We used The CASIA-IrisV3-Interval in our experiment because it is the only database for which we have access. The issue with this database is that there are some images which did not work with our system because they are corrupted, or because Mask segmentation rejects them. This mask segmentation problem is one which we faced in our experiment.

B. Modification of Image Quality

In our implementation, we added different noise to all images of CASIA-IrisV3-Interval to study the quality of images with those noises by following equation .

Noise image= Original image + numbers of noise (0.01,0.05,0.09 or 0.1) *randn(size(Original Image)) (see Figure 2)

Much testing was done to add noise with sub-data. We decide to add those number 0.01, 0.05,0.09 and 0.1 to all images on the database, because we found that they produced clearer results.

IV. EVALUATION OF THE IMPACT OF NOISE ON IRIS RECOGNITION

In our experiment, we used the Mask method which uses the Kovesei edge detector. Variation is known as Canny edge detector. The next step is applying the circular Hough transform to determine the iris/sclera that is correspondent to iris/pupil [8]

This Segmentation approach has the following in common:

- To detect the contrast between the pupil and the iris for the inner boundary.
- To detect the contrast between the sclera for the outer boundary, and the approximate distance between the boundaries of the circles around the iris.

After segmentation, an iris mask is generated.

Automatic segmentation is achieved using the circular Hough transformation method since it is able to localize the circular iris and pupil regions. The iris, pupil regions, and the linear Hough transform, are used for localizing the occluding eyelids. The Threshold was also used for isolating eyelashes and reflections. [8]

The next step, after the iris segmentation process, will be iris normalization. In order to allow comparisons, the iris region will transform into fixed dimensions. The purpose of this stage is to generate constant dimensions of the iris regions. The technique implemented is Daugman's Rubber sheet model [8], [14], [15], [18].

Features of the iris are encoded by contortion of the normalized iris region with 1D Log-Gabor filters and phase quantization of the output in order to generate a bitwise biometric template (see Figure 3 and 4).

The Hamming distance (HD) was selected as a matching metric that gave a measure of how many bits conflicted between the two templates [8]. Failure of statistical independence between two templates would result in a match; that is, the two templates were deemed to have

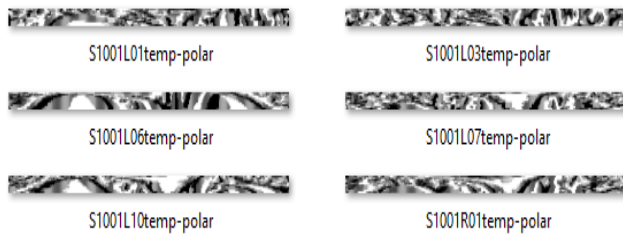


Figure 3: Samples template of iris with noise.



Figure 4: Samples mask of iris with noise.

been created from the same iris if the Hamming distance produced was lower than a constant Hamming distance [8].

When we got the HD results for each Noise data image, we built the ROC curve for each one according to True positive rates (TPR), and False positive rates (FPR), following these equations:

$$TPR = TP \div P = TP \div (TP + FN)$$

$$FPR = FP \div N = FP \div (FP + TN)$$

V. RESULTS

The higher level of noise is noted to provide less clearer images indicated by the values of the true positive rate. Consequently there appears to be an inverse relationship between the level of generated noise and quality of images. The ROC curve with 0.1 noise produced the lowest TPR value. The ROC curve with 0.01 showed the highest TPR value. The ROC curve with 0.05 has produced an intermediate value with regards to TPR. Images with 0.09 noise produced TRP value slightly better than that of 0.1 noise. The ROC curves demonstrated (Figure 5) that application of different levels of noise influenced the image quality. Images without noise have been shown to produce the clearest results with regards to quality of images in CASIA V3 database.

With noise the iris images are giving us significant reliability of images that is still good enough for identification.

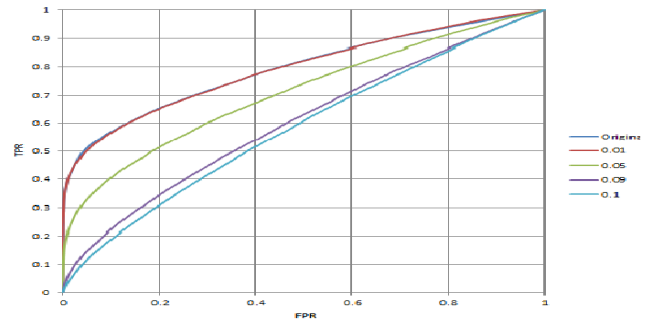


Figure 5: Comparison of the accuracy result on the CASIA datasets using leave-one-out cross validation.

VI. CONCLUSION

The iris of a human eye is the most suitable biometric trait to be used in the authentication system, because it provides a unique identification parameter.

The purpose of this paper is to prove that the reliability of iris is superior to other biometrics even in the presence of noise that could come with captured iris images.

There are some issues which need to be addressed in our work. First, we faced challenges with iris augmentations in which some of the images were not working with our system because of corruption of the images. Also, the automatic segmentation is insufficient, because it could not successfully segment the iris regions for all of the eye images.

However, there are many issues in biometric systems such as FMR and FNMR, which can affect the performance of biometric systems, especially if one of them has a high rate. These days, iris recognition systems have been undergone multiple studies and research to increase performance and minimize cost. Iris Recognition is one of the most important and preferable traits used for the identification and authentication of humans for different purposes [33]. The iris has a unique identity for each person. Our system was studying and implementing the various iris recognition schemes available by using Chinese academy of sciences-institute of automation (CASIA) database [7] to know the image quality of each image in CASIA V3 database.

The major advantage of the iris, is that when compared to other visual recognition techniques, we find that there is a huge degree of variability in the patterns between individuals. This meant that the large databases can be searched without resulting in any false matches [7].

Although the technique was introduced with a high level of security, there was still some chance of compromise. This meant that in spite of high security, there were chances to recover data, which may lead to the unwanted extraction of data. There are other techniques which could be used to make biometrics more secure, called biometric

cryptosystems. The use of Iris Biometric Cryptosystem is one of the most important and successful traits used for the authentication and identification of people for different legal or other purposes [33].

REFERENCES

- [1] M. Boyd, D. Carmaciu, F. Giannaros, T. Payne, W. Snell, and D. Gillies, "Iris recognition," Technical report, Department of Computer Science, Imperial College London, Tech. Rep., 2010.
- [2] A. Lin, S. Lin, and V. Yen, "Eye know you."
- [3] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Electronics in Marine, 2004. Proceedings Elmar 2004. 46th International Symposium*. IEEE, 2004, pp. 184–193.
- [4] A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross, "Biometrics: a grand challenge," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2. IEEE, 2004, pp. 935–942.
- [5] B. Dorizzi, "Biometrics at the frontiers, assessing the impact on society, technical impact of biometrics," *European Parliament Committee on Citizens Freedoms and Rights, Justice and Home Affairs (LIBE), Technical Report*, 2005.
- [6] E. S. Dunstone, "Emerging biometric developments: Identifying the missing pieces for industry," in *Signal Processing and its Applications, Sixth International Symposium on*. 2001, vol. 1. IEEE, 2001, pp. 351–354.
- [7] L. Li, F. Xu, H. Wang, C. She, and Z. Fan, "Chinese academy of sciences," 2004.
- [8] L. Masek *et al.*, "Recognition of human iris patterns for biometric identification," Ph.D. dissertation, Master's thesis, University of Western Australia, 2003.
- [9] S. Angle, R. Bhagtani, and H. Chheda, "Biometrics: A further echelon of security," in *UAE International Conference on Biological and Medical Physics*, 2005.
- [10] C. Rathgeb, "Iris-based biometric cryptosystems," Ph.D. dissertation, Diplomarbeit, Salzburg University, 2008.
- [11] A. Ross, A. Jain, and S. Pankati, "A prototype hand geometry-based verification system," in *Proceedings of 2nd Conference on Audio and Video Based Biometric Person Authentication*, 1999, pp. 166–171.
- [12] A. Ross and R. Govindarajan, "Feature level fusion in biometric systems," see <http://www.wvu.edu/~bknc/2004%20Abstracts/Feature%20Level%20Fusion%20in%20Biometric%20Systems.pdf>, 2004.
- [13] J. Daugman, "How iris recognition works," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 1, pp. 21–30, 2004.
- [14] J. G. Daugman, "Biometric personal identification system based on iris analysis," Mar. 1 1994, uS Patent 5,291,560.
- [15] P. Verma, M. Dubey, P. Verma, and S. Basu, "Daughman's algorithm method for iris recognition-a biometric approach," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 6, pp. 177–185, 2012.
- [16] P. D. Kovesi, "Matlab and octave functions for computer vision and image processing," Online: <http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/#match>, 2000.
- [17] K. Nguyen, C. Fookes, and S. Sridharan, "Fusing shrinking and expanding active contour models for robust iris segmentation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 185–188.
- [18] T. W. Hsiung and S. S. Mohamed, "Performance of iris recognition using low resolution iris image for attendance monitoring," in *Computer Applications and Industrial Electronics (ICCAIE), 2011 IEEE International Conference on*. IEEE, 2011, pp. 612–617.
- [19] W. W. Boles and B. Boashash, "A human identification technique using images of the iris and wavelet transform," *Signal Processing, IEEE Transactions on*, vol. 46, no. 4, pp. 1185–1188, 1998.
- [20] R. P. Wildes, J. C. Asmuth, G. L. Green, S. C. Hsu, R. J. Kolczynski, J. Matey, and S. E. McBride, "A system for automated iris recognition," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*. IEEE, 1994, pp. 121–128.
- [21] H. Ali, M. Salami *et al.*, "Iris recognition system by using support vector machines," in *Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on*. IEEE, 2008, pp. 516–521.
- [22] F. Zhao, Q. Huang, and W. Gao, "Image matching by normalized cross-correlation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 2. IEEE, 2006, pp. II–II.
- [23] X.-D. Zhu, Q.-I. Cui, Y.-N. Liu, and X. Ming, "A quality evaluation method of iris images sequence based on wavelet coefficients in" region of interest"," in *Computer and Information Technology, International Conference on*. IEEE Computer Society, 2004, pp. 24–27.
- [24] Y. Chen, S. C. Dass, and A. K. Jain, "Localized iris image quality using 2-d wavelets," in *Advances in Biometrics*. Springer, 2005, pp. 373–381.
- [25] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Personal identification based on iris texture analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1519–1533, 2003.
- [26] M. Salganicoff and G. H. Zhang, "Method of measuring the focus of close-up images of eyes," Sep. 14 1999, uS Patent 5,953,440.
- [27] B. J. Kang and K. R. Park, "A study on iris image restoration," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 31–40.
- [28] N. D. Kalka, J. Zuo, N. A. Schmid, and B. Cukic, "Image quality assessment for iris biometric," in *Defense and Security Symposium*. International Society for Optics and Photonics, 2006, pp. 62 020D–62 020D.
- [29] H. Proenca and L. A. Alexandre, "Toward noncooperative iris recognition: A classification approach using multiple signatures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 607–612, 2007.
- [30] M. Vatsa, R. Singh, and A. Noore, "Improving iris recognition performance using segmentation, quality enhancement, match score fusion, and indexing," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 38, no. 4, pp. 1021–1035, 2008.
- [31] P. Li and H. Ma, "Iris recognition in non-ideal imaging conditions," *Pattern Recognition Letters*, vol. 33, no. 8, pp. 1012–1018, 2012.
- [32] R. Parashar and S. Joshi, "Comparative study of iris databases and ubiris database for iris recognition methods for non-cooperative environment," *International Journal of Engineering*, vol. 1, no. 5, 2012.
- [33] C. Rathgeb and A. Uhl, "The state-of-the-art in iris biometric cryptosystems," *State of the art in Biometrics*, pp. 179–202, 2011.

Lucas-Kanade Scale Invariant Feature Transform for Uncontrolled Viewpoint Face Recognition

Yongbin Gao¹, Hyo Jong Lee^{1,2}

¹Division of Computer Science and Engineering,

²Center for Advanced Image and Information Technology
Chonbuk National University, Jeonju 561-756, Korea

Abstract - Face recognition has been widely investigated in the last decade. However, real world application for face recognition is still a challenge. Most of these face recognition algorithms are under controlled settings, such as limited viewpoint and illumination changes. In this paper, we focus on face recognition which tolerates large viewpoint change. A novel framework named Lucas-Kanade Scale Invariant Feature Transform (LK-SIFT) is proposed. LK-SIFT is an extension of SIFT algorithm. SIFT is a scale and rotation invariant algorithm, which is powerful for small viewpoint changes in face recognition, but it fails when large viewpoint change exists. To handle this problem, we propose to use Lucas-Kanade algorithm to generate different viewpoint face from a single frontal face. After that, SIFT is used to detect local features from these viewpoints, these SIFT features contain information of different viewpoint face, which can deal with the problem of face viewpoint change. Finally, our framework is compared with the SIFT algorithm and other similar solutions. Experiment results show our framework achieves better recognition accuracy than SIFT algorithm at the cost of acceptable computational time gains compared with other similar algorithms.

Keywords: Face recognition, Lucas-Kanade, Scale Invariant Feature Transform.

1 Introduction

Real world face recognition has many useful applications, such as identifying subjects from surveillance camera for public security and annotating people from digital photos automatically for individuals. There are some successful commercial face recognition systems available like Google Picasa and Apple iPhoto. However, face recognition research is still far from mature [1]. Earlier face recognition algorithms are only effective under controlled settings, such as the probe and gallery images are frontal. This algorithm fails when it is applied to cases as pose and illumination changes. This paper focuses on the viewpoint invariant face recognition, which identify face when probe faces are from different viewpoints while gallery faces are frontal.

The problem of face recognition under different viewpoint is the distance between different poses is bigger than distance between different subjects. One solution is to eliminate the distance between different poses. Among which,

face normalization is an effective method to remove the pose difference. Face normalization can be used as 2D or 3D model. As for 2D model, Markov Random Fields (MRF) is widely used to find correspondences between frontal face and the profile probe faces [2, 3]. MRF is to find 2D displacement by minimizing the energy, which consists of two parts, one is distance of corresponding node, another one represents the smoothness between neighbour nodes. Lucas-Kanade method is also used for face alignment [4, 5]. As for 3D model, Blanz et al. proposes an effective 3D morphable method to fit the 3D model to 2D face [6], the fitting shape and texture coefficients are used for face recognition. Normalization method can be used to construct the frontal face from the probe profile face [2]. It can also be used to directly match between a probe image and a gallery image and the matching score represent the similarity between two faces [3]. These normalization methods are effective at the cost of long computation time. It is reported that two minutes is needed to normalize one face [2]. Marsico et al. proposes a FACE framework to recognize face for uncontrolled pose and illumination changes [7]. It detects some keypoints using STASM algorithm [8], and construct half face by the middle line keypoints, the rest half face is reflected from the constructed half face. This easy method is fast but not robust for it highly depends on the accuracy of keypoints detection, when the keypoints detection fails, the system performance becomes bad.

A new classifier or new feature is proposed to deal with the viewpoint change problem. For the new classifier, one shot similarity (OSS) or two shot similarity (TSS) are proposed by introducing another dataset, which contains no probe and gallery images [9]. Each dataset contains different images of a single subject or different subjects viewed from a single pose. Similarity scores between two faces are calculated by the model built by one of faces and the introduced dataset using LDA or SVM. Cross-pose face recognition shares similar idea by introducing a third dataset [10]. Faces from different viewpoints are all linearly represented by the introduced dataset using subspace method, similarity between these faces is then calculated indirectly by the linear coefficients. As for new feature extraction, tied factor analysis is proposed to estimate the linear transformation and noise parameters in "identity" space [11].

Besides the exploration on face recognition, there are many researches on local descriptor, which is effective to deal

with affine transformation between two images. Such as Harris-Affine [12], Hessian-Affine [13], Affine SIFT [14] algorithms. These algorithms are powerful for planar object comparison, while human face is non-planar, which contains significant 3D depth. Directly using these algorithms don't work well, we propose LK-SIFT framework to deal with large viewpoint change for face recognition. We use Lucas-Kanade to generate a series of different viewpoint faces from a single frontal face. After that, SIFT is used to detect local features among all these viewpoints. Through our method, SIFT features contains enough information from all viewpoints to handle face pose variance.

The rest of this paper is organized as follows. Section II reviews the SIFT algorithm. We describe the proposed LK-SIFT framework in Section III. This algorithm includes image to image alignment and LK-SIFT algorithm. Section IV applies the above algorithm to FERET database, and presents the experiment results. Finally, we conclude this paper with future work in Section V.

2 Related work

Local features are effective methods for matching and recognition for it is robust to occlusion, scale, rotation or even affine transformation to some extent. Among these algorithms, Scale Invariant Feature Transform (SIFT) is a scale, rotation invariant local feature. It transforms image data into scale-invariant coordinates and localize the keypoint. Each keypoint is assigned a descriptor. The major steps for SIFT algorithm are as following [15]:

(1) Scale-space extrema detection: Image is transformed into different scales and size. Extrema are searched by finding maxima and minima over all scales using a difference-of-Gaussian scheme, which are invariant to scale and orientation.

(2) Keypoint localization: Extrema are refined by excluding poor localized or low contrast points by checking the refined location, scale and ratio of principal curvatures. This increases stability of keypoint localization.

(3) Orientation assignment: Each keypoint is assigned to one or more orientations based on local image gradient histogram. To provide scale and rotation invariance, local image data is transformed to the corresponding orientation and scale.

(4) Keypoint descriptor: Local keypoint descriptor is calculated around each keypoint by histogram of gradients. The descriptor is transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

There are several methods reported for image matching and recognition of SIFT algorithm, such as BBF [16], Hough transform [17]. Nearest neighbour is the original and effective matching method for SIFT features. SIFT features are first pre-extracted from gallery images and stored in a database. When matching with a probe image, each SIFT feature from the probe image is compared with all gallery features in database. Nearest neighbour and second nearest neighbour are searched based on the Euclidean distance. The ratio of these

two distances is compared with a threshold. Ratio that is smaller than the threshold is considered as a matching face.

The SIFT is scale and rotation invariant feature, but it is not affine invariant. Affine SIFT is the extension of SIFT algorithm. There are several parameters for affine transformation as:

$$A = H_\lambda R_1(\psi) T_t R_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \quad (1)$$

where λ , R_1 and T_t are a scale parameter, rotated angle, and tilted angle, respectively. Fig. 1 shows the geometric interpretation of these parameters. SIFT algorithm is just scale (λ) and rotation (ψ) invariant. The left t and ϕ are not invariant, Therefore, SIFT algorithm is not fully affine invariant. Affine SIFT is trying to fulfil the t and ϕ invariant.

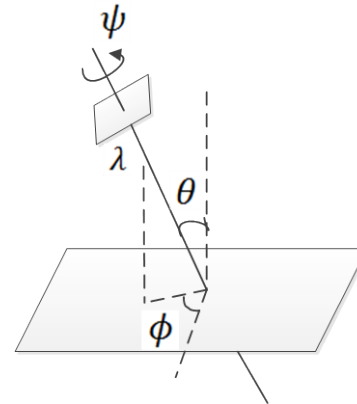


Fig. 1 Geometric interpretation of affine decomposition. λ and ψ are scale and rotation from camera. θ and ϕ is tilt and rotation of subject, which named latitude and longitude respectively. Where $t = 1/\cos\theta$.

Affine SIFT transforms an image into a series of simulated images by the change of longitude ϕ and latitude θ [14]. These simulated images are sampled to achieve a balance between accuracy and sparsity. However, Affine SIFT generates 61 images when the number of tilts set to 7. This increases the computation time too much, which is also unnecessary for face recognition. Moreover, human face contains 3D depth, while affine transformation is effective for planar object, simple affine transformation for a holistic face is not enough to represent the pose variant of face. In this paper, we propose to use LK-SIFT algorithm to simulate different pose from a single frontal face.

3 LK-SIFT

3.1 Image to Image Alignment

Lucas-Kanade algorithm is first used as an effective image alignment method [18]. Image alignment is to find correspondences between gallery and probe images, firstly we equally divide image into several subregions, for pixels in the same subregions, we assume they share the same warp parameters, Let the warp function be $x' = W(x, P)$, where $P = [p_1, p_2, \dots, p_m]^T$, For affine warp, $m=6$, and

$$W(x, P) = \begin{pmatrix} 1 + p_1 & p_3 & p_5 \\ p_2 & 1 + p_4 & p_6 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (2)$$

Fig. 2 shows two images captured at two different poses, where I represents the probe image and T represents the gallery image. We divide image T into non-overlap subregions with same size. For each subregion r in T , we try to find a warp that aligns these two images. I_r is the corresponding subregion to T_r after warp transformation. The main objective for alignment is to minimize the error between the T_r and the warped subregions I_r as:

$$E_r = \sum_x (I_r(W(x, P)) - T_r(x))^2 \quad (3)$$

The solution for Equation 3 is to iterate calculating a ΔP and update P till P converge. Lucas-kanade gives a solution for calculating ΔP by:

$$\Delta P = H_{img}^{-1} \sum_x \left(\nabla I_r \frac{\partial W}{\partial P} \right)^T (T_r(x) - I_r(W(x, p))) \quad (4)$$

where $\nabla I_r = \left(\frac{\partial I_r}{\partial x}, \frac{\partial I_r}{\partial y} \right)$ is the gradient of I_r . $\frac{\partial W}{\partial P}$ is the Jacobian of the warp (shown in Eq. 2). H_{img} is the pseudo Hessian matrix, which is given by:

$$H_{img} = \sum_x \left(\nabla I_r \frac{\partial W}{\partial P} \right)^T \left(\nabla I_r \frac{\partial W}{\partial P} \right) \quad (5)$$

We can now update the warp parameters $P \leftarrow P + \Delta P$ and iterate till the parameters P converge. This procedure is applied independently for every patch/subregion.

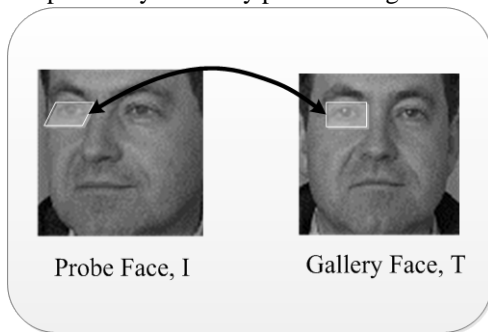


Fig. 2 Image to image alignment, image is divided into several subregions T_r , a warp between two subregions T_r and I_r is calculated by minimizing the alignment error.

3.2 LK-SIFT framework

Image to image alignment can be used to online recognition between two images. There are two kinds of online recognition methods. The first one is to calculate a match score for two images based on the warp parameters or alignment errors. Another one is to normalize images by transforming the profile face to its frontal face. However, these online alignments require long computational time, which is not good for real time applications. Another scheme is off-line alignment. Warps parameters are trained from several stack images, each stack images are from the same pose. There are two strategies for off-line alignment. First one is to average two set of stack images, and learn the warp parameters between two average images. Another strategy is to find warp parameters that minimize all images from the two set of stack images as [5]:

$$E_{r(stk)} = \sum_j \sum_x (I_{j,r}(W(x, P)) - T_{j,r}(x))^2 \quad (6)$$

where $I_{j,r}$ and $T_{j,r}$ are the r -th subregion of the j -th image from two set of stack images. The solution of Eq. 6 for ΔP is then as:

$$\Delta P = H_{(stk)}^{-1} \sum_j \sum_x \left(\nabla I_{j,r} \frac{\partial W}{\partial P} \right)^T (T_{j,r}(x) - I_{j,r}(W(x, p))) \quad (7)$$

Let $\Omega = [P_1, P_2, \dots, P_N]$ be the warp parameters for all the subregions between two viewpoints. Then we can learn a series of Ω for different viewpoints from a single frontal view as shown in Fig. 3. It is wise to generate viewpoint from nearby pose to reduce the alignment error. In our experiment, each pose is generated one by one from a slight pose change to larger one.

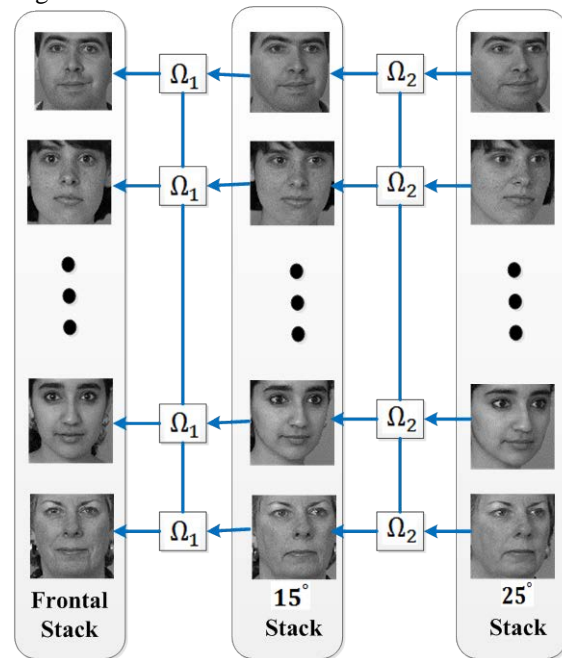


Fig. 3 Offline warps learning of different poses from frontal face. The warps are learned from nearby stacks.

In general, our proposed LK-SIFT can be summarized as following:

LK-SIFT: Lucas-Kanade Scale Invariant Feature Transformation

Pre-computed:

1. Learn a set of warp parameters $\Omega_1, \Omega_2, \dots, \Omega_N$ of N different poses from frontal face using N stack images, each stack images are from the same pose as shown in Fig. 3.
2. For each gallery (frontal) face, we generate N poses using the learned sets of warp parameters.
3. Compute the SIFT keypoints of these N pose faces and stored as a keypoint database.

Recognition:

1. For each probe face, compute its SIFT keypoints and compare these keypoints with keypoints of each subject in the keypoint database.
2. The subject that has the maximum number of matching keypoints with the probe face is considered as recognized subject.

LK-SIFT computes SIFT keypoints from several points of view of the gallery face. It can handle the pose change of probe face to some extent. The number of pose generated from the gallery face should be chosen wisely to achieve a balance between accuracy and sparsity. The step between two nearby pose should not be either too big or small. A big step is not enough to guarantee the recognition accuracy, and a small step result in too many redundant keypoints, which increase the computational time. From our experiment, we conclude that pose change from -15 degree to 15 degree, SIFT can achieve high accuracy. Therefore, choose a step of 15 degree for nearby pose is reasonable to achieve a high accuracy.

4 Results

In our experiments, we used FERET [19] grey database to evaluate our algorithm. This database contains 200 subjects, each subject contains 9 images captured from different poses. For each subject, we use frontal image as gallery, and other 8 pose images as probe images, the pose angle of which are -60, -40, -25, -15, 15, 25, 40 and 60 degrees, respectively. Figure 4 shows the different pose generated from frontal face using LK algorithm. First and third lines are the original database images of different pose; second and fourth lines are the generated pose faces using learned warps.



Figure 4. Different poses generated from frontal face using LK algorithm. First and third lines are the original database images of different pose; second and fourth lines are the generated pose faces using learned warps.

The parameters used in our experiment for SIFT algorithm are: image is resized to resolution of 200*200, and the ratio of nearest neighbour for SIFT is set to 0.8. Table I shows the comparison experiment results of recognition with ASIFT [14] and SIFT method. The number of tilt for ASIFT

is set to 3. The number of tilt means the affine transformation times for θ or t . When it sets to 3, ASIFT generates 10 viewpoints. For LK-SIFT algorithm, it generates 9 poses. From the table, we know that SIFT can get good results when a pose degree is between -15 to 15 degree, but LK-SIFT achieves better results than SIFT, ASIFT, especially under large pose change.

TABLE I EXPERIMENT RESULTS TO RECOGNIZE FACE WITH DIFFERENT POSE ON FERET DATABASE

Pose (°)	SIFT (%)	ASIFT(%)	LK-SIFT(%)
-40	53	56	63.5
-25	96.5	92.5	95.5
-15	99.5	99.5	98
15	100	99	98
25	92	93	93.5
40	48	58	59.5
Average	82	83	85

5 Conclusions

In this paper, a novel framework named Lucas-Kanade Scale Invariant Feature Transform (LK-SIFT) is proposed. LK-SIFT is an extension of SIFT algorithm, which is scale and rotation invariant. SIFT algorithm is powerful for small viewpoint changes in face recognition, but it fails when large viewpoint change exists. To handle this problem, we propose to use Lucas-Kanade algorithm to generate different viewpoint faces from a single frontal face. After that, SIFT is used to detect local features from these viewpoints, these SIFT features contain information from different viewpoint faces, which can deal with the problem of face viewpoint change. Finally, our framework is compared with the SIFT algorithm and other similar solutions. Experiment results show SIFT can get good results when pose degree is between -15 to 15 degree, but LK-SIFT achieve better result than SIFT, ASIFT, especially under large pose different. The computation time for LK-SIFT is smaller than ASIFT.

Acknowledgement: This work (Grants No. C0112553) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2013. This work was also supported by the Brain Korea 21 PLUS project, National Research Foundation of Korea.

6 References

- [1] G. Hua, M. H. Yang, E. L. Miller, Y. Ma, M. Turk, D.J. Kriegman and T. S. Huang, "Introduction to the Special Section on Real- World Face Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 10, pp. 1921–1924, Oct. 2011.

- [2] H. T. Ho, R.Chellappa, "Pose-Invariant Face Recognition Using Markov Random Fields," *IEEE Trans. Image Processing*, vol.22, no.4, pp.1573-1584, Apr. 2013.
- [3] S. R. Arashloo and J. Kittler, "Energy Normalization for Pose- Invariant Face Recognition Based on MRF Model Image Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no.6, pp. 1274-1280, June. 2011.
- [4] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no.3, pp.221 – 255, Mar. 2004.
- [5] A. B. Ashraf, S. Lucey and T. Chen, "Learning Patch Correspondences for Improved Viewpoints Invariant Face Recognition," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2008, pp. 1-8.
- [6] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1-12, sep. 2003.
- [7] M. D. Marsico, M. N. D. Riccio and H. Wechsler, "Robust Face Recognition for Uncontrolled Pose and Illumination Changes," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 43, no. 1, pp. 149-162, Jan. 2013.
- [8] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 504–513.
- [9] L. Wolf, T. hassner, and Y. Taigman, " Effective Unconstrained Face recognition by Combining Multiple Descriptors and Learned Background Statistics," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1978-1990, Oct. 2011.
- [10] A. Li, S. Shan, and W. Gao, "Coupled Bias–Variance Tradeoff for Cross-Pose Face Recognition," *IEEE Trans. Image Processing*, vol. 21, no. 1, pp. 305-315, Jan. 2012.
- [11] S. J. D. Prince, J. H. Elder, j. Warrell, and Fatima, "Tied Factor Analysis for Face Recognition across Large Pose Differences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 970-982, June. 2008.
- [12] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," *Int'l J. Computer Vision*, vol. 1, no. 60, pp. 3-86, 2004.
- [13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L.V. Gool, "A Comparison of Affine Region Detectors," *Int'l J. Computer Vision*, vol. 65, no. 1-2, Nov. 2005.
- [14] J. M. Morel and G. Yu, "ASIFT, A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp.438-469, 2009.
- [15] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, Nov. 2004.
- [16] J. Beis, and D. G. Lowe, "Shape indexing using approximate nearest-neighbour search in highdimensional spaces," In *Proc. Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 1000-1006.
- [17] Hough, P.V.C. 1962. Method and means for recognizing complex patterns. U.S. Patent 3069654.
- [18] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," In *Proc. International Joint Conference on Artificial Intelligence*, 1981, vol. 2, pp. 674–679.
- [19] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

Effective Kernel Mapping for One-Dimensional Principal Component Analysis in Finger Vein Recognition

Sepehr Damavandinejadmonfared, Vijay Varadharajan

Advanced Cyber Security Research Centre
Dept. of Computing, Macquarie University
Sydney, Australia

Abstract - Kernel functions have been very useful in data classification for the purpose of identification and verification so far. Applying such mappings first and using some methods on the mapped data such as Principal Component Analysis has been proven novel in many different areas. A lot of improvements have been proposed on PCA such as Kernel Principal Component Analysis, and Kernel Entropy Component Analysis which are known as very novel and reliable methods in face recognition and data classification. In this paper, we implemented four different Kernel mapping functions on finger database to determine the most appropriate one in terms of analyzing finger vein data using 1D-PCA. Extensive experiments have been conducted for this purpose using Polynomial, Gaussian, Exponential and Laplacian Principal Component Analysis (PCA) in 4 different examinations to determine the most significant one.

Keywords: Biometrics, finger vein recognition, Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA).

1 Introduction

The importance of reliability in verification and identification has gained lots of attention recently[1]. Finger vein is a newly proposed method of biometrics which has been able to gain many researchers' attention due to the fact that it is something internal and reliable to be used for this purpose. Furthermore, it has been proven by the medical studies that finger-vein pattern is unique and stable[2]. As the data in finger vein recognition[3][4][5] is 'image', some face recognition algorithms[6][7][8][9][10] have been proposed to be used in this case. Principal Component Analysis (PCA)[11][7][6] is one of the common and known methods of pattern recognition and face recognition[12][13] which has been used a lot in biometrics. PCA, however, is a linear method which makes it unable to properly deal with nonlinear patterns which might be in data. To overcome the mentioned drawback of PCA, Kernel Principal Component Analysis[14][8] was proposed, which is known to be more

appropriate than Principal Component Analysis in many cases such as pattern recognition and face recognition. It is because of the fact that using kernel function in the system makes it nonlinear. The mentioned reasons motivated us to conduct a comparative analysis between two known and mostly used methods called Principal Component Analysis (PCA) and Kernel Principle Component Analysis (KPCA)[9][15][16][17] in finger vein recognition. The main difference between PCA and KPCA is the fact that PCA is a linear method, while KPCA is the nonlinear version of PCA in which Kernel transforming is used. In PCA it is ensured that the transferred data is uncorrelated, and only preserve maximally the second-order statistics of the original data, which is why PCA is known as insensitive to the dependencies of multiple features of the pattern. In KPCA the mentioned problem has been overcome as it is not a linear method. In kernel Principal Component Analysis, however, it is essential that which kernel mapping function is chosen to be used. It could be considered very important due to the fact that each kernel mapping has particular characteristics and the data after being mapped will be in a totally different and high dimensional space where it could be too complicated to extract the valuable features. As Principal Component Analysis is a well-known method of dimensionality reduction, the combination of PCA and kernel mapping will lead to a more reliable system. There are several different types of kernel mapping which have been proven to be novel in different machine learning algorithms. In this research, we use four famous kernel mappings such as Polynomial, Gaussian, Exponential and Laplacian as they have an extensive use within image processing related algorithms. Comparison in this paper is between Different types of KPCA using the mentioned four kernel functions to map the data to achieve a two-fold contribution; first: Kernel Principal Component analysis is appropriate enough to be used in finger vein area, and second: which kernel mapping function is the most superior one.

The remaining of this paper is organized as follows:

In Section 2, Image acquisition is explained. In Section 3, Principal Component Analysis is explained. In Section 4, Kernel Principal Component Analysis (KPCA) is introduced. In section 5, experimental results on the finger vein database are given. Finally, section 6 concludes the paper.

2 Image Acquisition

Based on the proven scientific fact that the light rays can be absorbed by deoxygenated hemoglobin in the vein, absorption coefficient (AC) of the vein is higher than other parts of finger. In order to provide the finger vein images, four low cost prototype devices are needed such as an infrared LED and its control circuit (Osram SFH485 infrared light emitting diodes (IR LED) with wavelength 880nm which is located at the top of the design), a camera to capture the images (Logitech V-UAV35 web-cam is employed as the capturing devices at the bottom of our design), a micro-computer unit (MCU) to control the LED array, and a computer to process the images. The web-cam has an IR blocking filter; hence, it is not sensitive to the infrared (IR) rays. To solve this problem a negative film is used instead of IR blocking filter to prevent the infrared rays from being blocked. The negative film can operate as IR pass filter to transmit about 90% of radiation wavelength of 850nm [2]. Figure 1 shows an example of samples captured and cropped for this research.

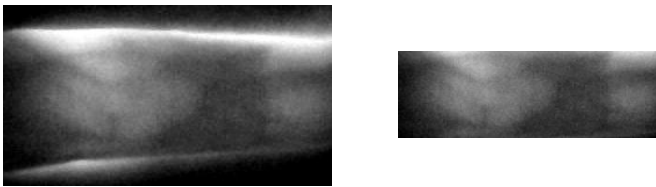


Fig. 1 : Original and cropped image

3 Principal Component Analysis (PCA)

Principal component analysis is known as a very powerful method for feature extraction. The usage of extracting eigenvectors and their corresponding eigenvalues to project the input data onto has been very common in image analysis such as face recognition and image classification. PCA, actually, extracts the features from the data and reduces the dimension of it. When the features are extracted, a classifier can be applied to classify them and the final decision can be made. Euclidian distance is used in our

algorithm which is very fast and sufficient to our purpose. In the rest of this section, PCA is explained briefly:

First the mean center of the images is computed. m represents the mean image.

$$m = \frac{1}{M} \sum_{i=1}^M X_i \quad (1)$$

The mean centered image is calculated by Eq. (2)

$$w_i = X_i - m \quad (2)$$

First covariance matrix is calculated by:

$$C = WW^T \quad (3)$$

Where W is a matrix composed of the column vectors w_i placed side by side.

Assuming that λ is eigenvector and v is eigenvalue, by solving $\lambda v = Cv$ eigenvectors and eigenvalues could be obtained.

By multiplying both side by W and substitution of C we can get the following equation.

$$WW^T(Wv) = \lambda(Wv) \quad (4)$$

Which means the first $M-1$ eigenvectors λ and eigenvalues v can be obtained by calculating WW^T .

When we have M eigenvectors and eigenvalues the images could be projected onto $L \ll M$ dimensions by computing

$$\Omega = [v_1 v_2 \dots v_L]^T \quad (5)$$

Where, Ω is the projected value. Finally, to determine which face provides the best description of an input image the Euclidean distance is calculated using equation (6).

$$\epsilon_k = \|\Omega - \Omega_k\| \quad (6)$$

And finally, the minimum ϵ_k will decide the unknown data into k class.

4 Kernel Principal Component Analysis (KPCA)

Unlike PCA, KPCA extracts the features of the data nonlinearly. It obtains the principal components in F which is a high dimensional feature space that is related to the feature spaces nonlinearly. The main idea of KPCA is to

map the input data to the feature space F first using a nonlinear mapping Φ . when input data have nonlinearly been mapped, the Principle Component Analysis (PCA) will be performed on the mapped data [3]. Assuming that F is centered, $\sum_{i=1}^M \Phi(X_i) = 0$ where M is the number of input data. The covariance matrix of F can be defined as

$$C = \frac{1}{M} \sum_{i=1}^M \Phi(X_i) \cdot \Phi(X_i)^T \quad (7)$$

To do this, this equation $\lambda v = C v$ which is the eigenvalue equation should be solved for eigenvalues $\lambda \geq 0$ and eigenvectors $v \in F$.

As $Cv = (1/M) \sum_{i=1}^M (\Phi(X_i) \cdot v) \Phi(X_i)$, solutions for v with $\lambda \neq 0$ lie within the span of $\Phi(X_1), \dots, \Phi(X_M)$, these coefficients $\alpha_i (i = 1, \dots, M)$ are obtained such that

$$V = \sum_{i=1}^M \alpha_i \Phi(X_i) \quad (8)$$

The equations can be considered as follows

$$\lambda(\Phi(X_i) \cdot V) = (\Phi(X_i) \cdot Cv) \quad \text{for all } i = 1, \dots, M \quad (9)$$

Having $M \times M$ matrix K by $K_{ij} = k(X_i, X_j) = (\Phi(X_i) \cdot \Phi(X_j))$, causes an eigenvalue problem.

The Solution to this is

$$M \lambda \alpha = K \alpha \quad (10)$$

By selecting the kernels properly, various mappings can be achieved. One of these mappings can be achieved by taking the d -order correlations, which is known as ARG, between the entries, X_i , of the input vector X . The required computation is prohibitive when $d > 2$.

$$\begin{aligned} (\Phi_d(X) \cdot \Phi_d(y)) &= \sum_{i_1, \dots, i_d=1}^N x_{i_1} \dots x_{i_d} \cdot y_{i_1} \dots y_{i_d} \\ &= \\ (\sum_{i=1}^N x_i \cdot y_i)^d &= (x \cdot y)^d \end{aligned} \quad (11)$$

To map the input data into the feature space F , there are four common methods such as linear (polynomial degree 1), polynomial, Gaussian, and sigmoid, which all are examined in this work in addition to Principal Component Analysis.

5 Experimental Results

In this section, the experiments are conducted to corroborate the performance of Gaussian Kernel Principal Component Analysis (KPCA) over other kinds of KPCA such as Polynomial, Exponential, and Laplacian PCA in terms of finger vein recognition. Finger vein database used in the experiments consists of 500 images from 50 individuals; 10 samples from each subject were taken. In this experiment 4, 5, 6 and 7 images are used to train and the remaining 6, 5, 4 and 3 images are used to test respectively. In each experiment, the accuracy is calculated using the first 100 components of the extracted features meaning that each experiment is repeated 100 times using the first 100 features to project the data onto, and also the dimension is reduced from 60% to 85% in different experiments. The results are shown in Figure 2, 3, 4 and 5.

As it was expected, use of kernel functions to map the data first and then applying PCA on the mapped data (KPCA) results in acceptable accuracies varying from over 70% up to near 100% in different experiments. Polynomial KPCA, however, seems to be the worst among all types of KPCA and there is a great discrepancy between Polynomial and other kinds of kernel KPCA in terms of final outputs of the system. The results show that Polynomial kernel reaches its optimized point when using even less than 10 components and it remains the same no matter how many more components to be used. It could be considered an advantage as using this kernel can be faster than others as it gets to its peak in the point 10 or less than that. The accuracy in Polynomial KPCA, however, is not satisfying at all and is less than the others in almost all experiments. From another point of view, when 4 images are used to train, the highest accuracy obtained is around 95 percent while the accuracy rate almost reaches 99 when using 7 images to train which means, the more the number of training images is, the higher accuracy gets.

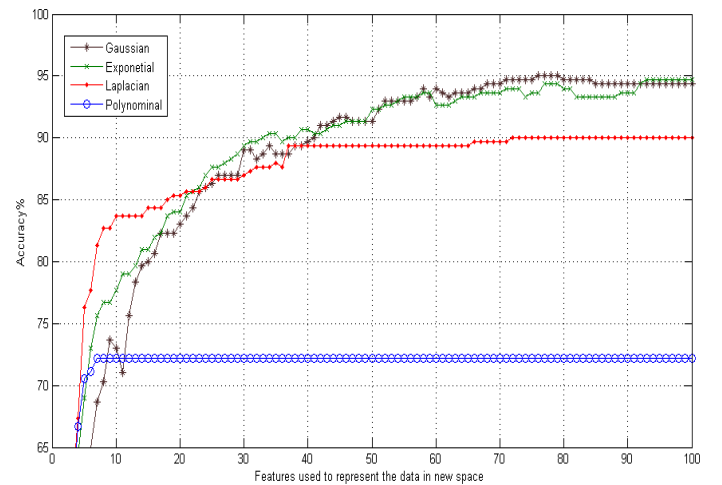


Fig. 2 : Comparison of accuracies obtained using 4 images to train and 6 to test

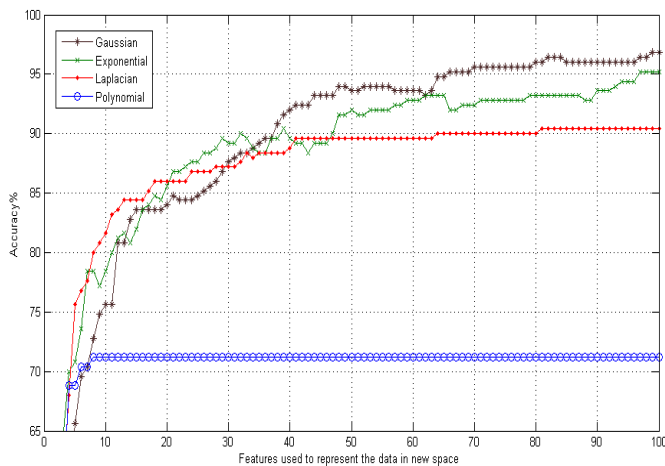


Fig. 3 : Comparison of accuracies obtained using 5 images to train and 5 to test

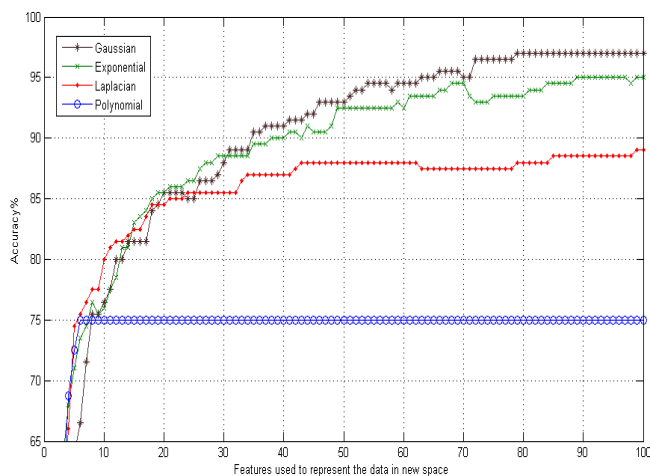


Fig. 4 : Comparison of accuracies obtained using 6 images to train and 4 to test

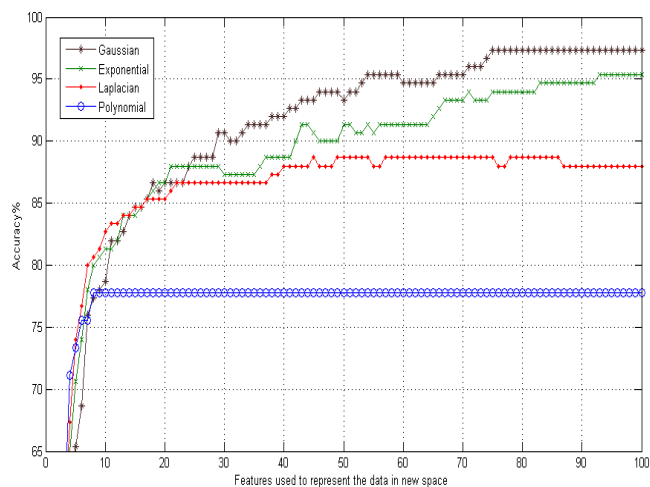


Fig. 5 : Comparison of accuracies obtained using 7 images to train and 3 to test

It is observed from the results that the accuracies achieved using Laplacian KPCA are not very close to those of the Exponential and Gaussian methods; it is also understood that although the accuracies of Gaussian KPCA are close to those of Exponential KPCA, Exponential method in a majority of implementations results in less accuracy compared to Gaussian. Results indicate that in the first experiment, where 4 images were used to train and the remaining 6 images to test, the difference between the accuracies obtained using Gaussian, Exponential, and Laplacian were not that significant as Laplacian, Exponential, and Gaussian reached 90%, 94%, and 95% respectively. However, the more the number of training images get, the higher accuracy Gaussian KPCA obtains and its discrepancy in accuracy becomes larger to the point that leads to the conclusion that Gaussian Kernel Principal Component Analysis is the most superior kernel mapping in finger vein recognition systems

6 Conclusion

The performance of four different types of Kernel Principal Component Analysis on finger vein recognition have been validated in this paper. It is shown that not only is the Gaussian Kernel Principal Component Analysis the most appropriate one in comparison to the other types of KPCA (polynomial, Exponential, and Laplacian), but also this method is efficient enough to be used in finger vein recognition as for such a big data base the accuracy is very high and promising.

7 References

- [1] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Syst.*, vol. 14, no. 1, pp. 4–20, 2004.
- [2] A. Kumar and Y. Zhou, "Human identification using finger images.," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2228–44, Apr. 2012.
- [3] T. S. Beng and B. A. Rosdi, "Finger-vein identification using pattern map and principal component analysis," *2011 IEEE Int. Conf. Signal Image Process. Appl.*, pp. 530–534, Nov. 2011.
- [4] W. Song, T. Kim, H. C. Kim, J. H. Choi, H.-J. Kong, and S.-R. Lee, "A finger-vein verification system using mean curvature," *Pattern Recognit. Lett.*, vol. 32, no. 11, pp. 1541–1547, Aug. 2011.
- [5] J.-D. Wu and C.-T. Liu, "Finger-vein pattern identification using principal component analysis and the neural network technique," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5423–5427, May 2011.
- [6] C. Wen and J. Zhang, "Palmprint Recognition based on Gabor Wavelets and 2-Dimensional PCA&PCA," in *International Conference on Wavelet Analysis and Pattern Recognition*, 2007, pp. 2–4.

- [7] C. Vision, "A Comparison of Face Recognition Methods Final Project Report A Combined Project for," *Artif. Intell.*, 2003.
- [8] R. M. Ebied, "Feature Extraction using PCA and Kernel-PCA for Face Recognition," *Int. Conf. INFOrmatICS Syst.*, vol. 8, pp. 72–77, 2012.
- [9] V. D. M. Nhat and S. Lee, "Kernel-based 2DPCA for Face Recognition," *2007 IEEE Int. Symp. Signal Process. Inf. Technol.*, pp. 35–39, Dec. 2007.
- [10] C. Yu, H. Qing, and L. Zhang, "K2DPCA Plus 2DPCA: An Efficient Approach for Appearance Based Object Recognition," *2009 3rd Int. Conf. Bioinforma. Biomed. Eng.*, pp. 1–4, Jun. 2009.
- [11] S. Lin and D. Ph, "An Introduction to Face Recognition Technology," *Pattern Recognit.*, no. 1995, pp. 1–7, 1997.
- [12] W. Zuo, D. Zhang, and K. Wang, "Bidirectional PCA with assembled matrix distance metric for image recognition.," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 36, no. 4, pp. 863–72, Aug. 2006.
- [13] D. Zhang and Z.-H. Zhou, "Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, no. 1–3, pp. 224–231, Dec. 2005.
- [14] K. I. Kim, K. Jung, and H. J. Kim, "Face Recognition using Kernel Principal Component Analysis," *Signal Processing*, vol. 9, no. 2, pp. 40–42, 2002.
- [15] P. Hu and A. Yang, "Indefinite Kernel Entropy Component Analysis," *Sci. Technol.*, no. 3, pp. 0–3, 2010.
- [16] R. Jenssen, "Kernel entropy component analysis.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 847–60, May 2010.
- [17] B. H. Shekar, M. Sharmila Kumari, L. M. Mestetskiy, and N. F. Dyshkant, "Face recognition using kernel entropy component analysis," *Neurocomputing*, vol. 74, no. 6, pp. 1053–1057, Feb. 2011.

A Facial Expression Recognition System from Depth Video

Md. Zia Uddin

Department of Computer Education
Sungkyunkwan University
Seoul, Republic of Korea
Email: ziauddin@skku.edu

A. M. Jehad Sarkar

Department of Digital Information Engineering
Hankuk University of Foreign Studies
Yongin, Republic of Korea
Email: jehad@hufs.ac.kr

Abstract— In this work, a novel approach is proposed to recognize some facial expressions from time-sequential depth videos. Local Directional Pattern (LDP) features are extracted from the time-sequential depth faces that are followed by Linear Discriminant Analysis (LDA) to make the features more robust. Finally, the robust local features are applied with Hidden Markov Models (HMMs) for facial expressions successfully. The proposed approach shows superior recognition rate against the conventional approaches.

Keywords—Depth Information, LDP; HMM; FER

I. INTRODUCTION

Facial expression recognition (FER) provides machines a way of sensing a peoples' emotion that can be considered one of the mostly used artificial intelligence and pattern analysis applications [1]-[10]. In case of extracting expression images through RGB cameras, most of the FER works used Principal Component Analysis (PCA), which is very well known for dimension reduction and used in many earlier works. In [3], PCA was used to recognize Facial Action Units (FAUs) from the facial expression images. In [5] as well as [6], PCA was used for FER with the Facial Action Coding System (FACS). Very recently, Independent Component Analysis (ICA) has been extensively utilized for FER based on local face image features [5], [10], [11]-[21]. In [14], the authors used ICA to extract local features and then classified several facial expressions. In [15], ICA was used to recognize the FAUs. Besides ICA, Local Binary Patterns (LBP) has been used lately for FER [22]-[24]. The main property of LBP features is their tolerance against illumination changes as well as their computational simplicity. Later on, LBP was improved by focusing on face pixel's gradient information and named as Local Directional Pattern (LDP) to represent local face features [25]. As like as LBP, LDP features also have the tolerance against illumination changes but they represent much robust features than LBP due to considering the gradient information for each pixel as aforementioned [25]. Thus, LDP can be a robust approach and hence can be adopted for FER. To make LDP face features more robust, Linear Discriminant Analysis (LDA) can be applied as LDA is considered to be a

robust tool to be used to obtain good discrimination among the face images from different expressions by considering linear features spaces. Hidden Markov Model (HMM) is considered to be a robust tool to model and decode time-sequential events [21], [26]-[28]. Hence, HMM seems an appropriate choice to train and recognize features of different facial expressions for FER.

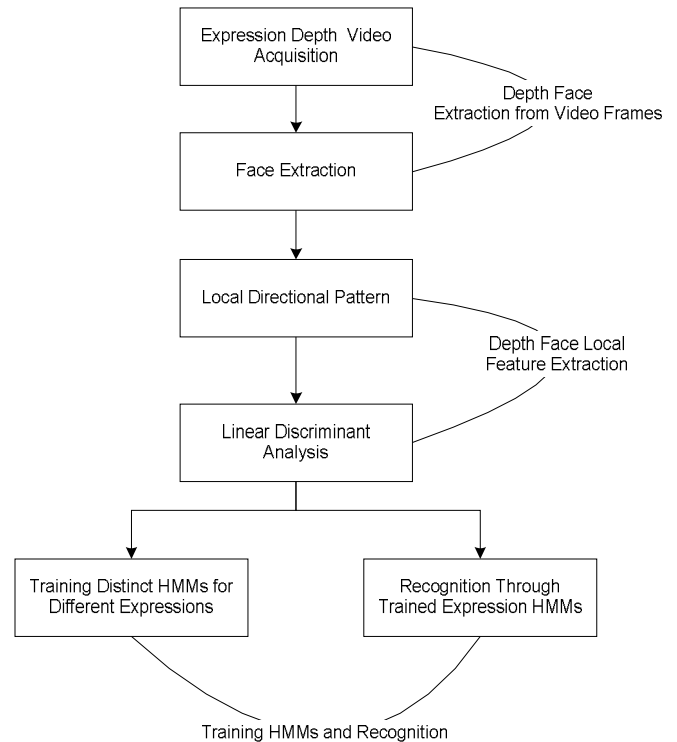


Fig. 1. Basic steps involved in the proposed facial expression recognition system.

For capturing face images, RGB cameras are used most widely but the faces captured through a RGB camera cannot provide the depth of the pixels based on the far and near parts of human face in the facial expression video where the depth information can be considered to contribute more to extract efficient features to describe the expression more strongly.

Hence, depth videos should allow one to come up with more efficient person independent FER.

II. PREPROCESSING

The images of different expressions are captured by a depth camera [29] where the camera generates RGB and distance information (i.e., depth) simultaneously for the objects captured by the camera. The depth video represents the range of every pixel in the scene as a gray level intensity (i.e., the longer ranged pixels have darker and shorter ones brighter values or vice versa). Fig. 1 shows the basic steps of proposed FER system.

Fig. 2(a) represents a depth image from a surprise expression. It can be noticed that in the depth image, the higher pixel value represents the near (e.g., nose) and the lower (e.g., eyes) the far distance. The pseudo-color image corresponding to the depth image in Fig. 2(b) also indicates the significant differences among different face portions where the color intensities. Fig. 3 shows five generalized depth faces from a disgust expression.

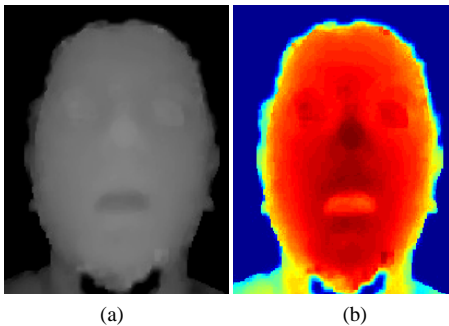


Fig. 2. (a) A depth image and (b) corresponding pseudo color image of a surprise image.

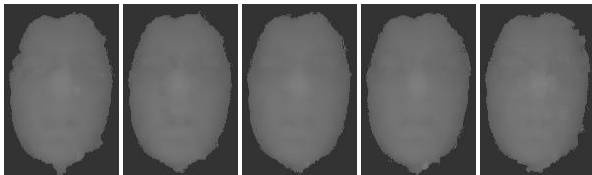


Fig. 3. A sequential depth facial expression images of disgust.

III. FEATURE EXTRACTION

The Local Directional Pattern (LDP) assigns an eight-bit binary code to each pixel of an input depth image. This pattern is then calculated by comparing the relative edge response values of a pixel in eight different directions. Kirsch, Prewitt and Sobel edge detector are some of the different representative edge detectors that can be used. Amongst which, the Kirsch edge detector [20] detects the edges more accurately than the others as it considers all eight neighbors. Given a central pixel in the image, the eight directional edge response values $\{m_k\}$, $k=0,1,\dots,7$ are computed by Kirsch masks M_k in eight different orientations centered on its

position [18]. Fig. 4 shows these masks.

The presence of a corner or an edge represents high response values in some particular directions and therefore, it is interesting to know the p most prominent directions in order to generate the LDP. Here, the top- p directional bit responses b_k are set to 1. The remaining bits of 8-bit LDP pattern are set to 0. Finally, the LDP code is derived by (1). Fig. 5 shows the mask response as well as LDP bit positions and Fig. 6 an exemplary LDP code with $d=4$.

$$LDP_p = \sum_{k=0}^7 B_k (m_k - m_p) \times 2^k, \quad B_k(a) = \begin{cases} 1 & a \geq 0 \\ 0 & a < 0 \end{cases} \quad (1)$$

where, m_p is the p -th most significant directional response.

$\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix}$	$\begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix}$	$\begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$	$\begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$
east M_0	north east M_1	north M_2	north west M_3
$\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix}$	$\begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix}$	$\begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix}$	$\begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$
west M_4	south west M_5	south M_6	south east M_7

Fig. 4. Kirsch edge masks in eight directions.

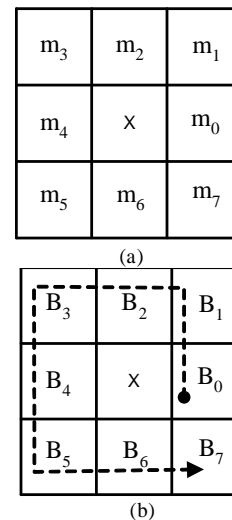


Fig. 5. (a) Edge response to eight directions and (b) LDP binary bit positions.

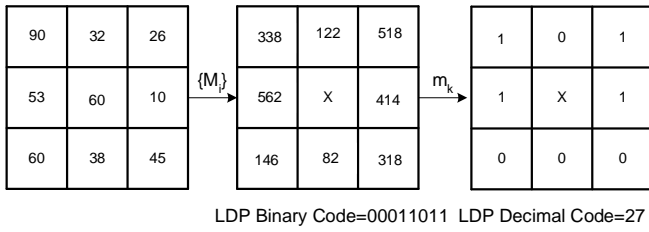


Fig. 6. LDP code.

Thus, an image is transformed to the LDP map using LDP code. The image textual feature is presented by the histogram of the LDP map of which the q^{th} bin can be defined as

$$T_q = \sum_{x,y} I \{LDP(x, y) = q\}, q = 0, 1, \dots, n-1 \quad (2)$$

where n is the number of the LDP histogram bins (normally $n = 256$) for an image I . Then, the histogram of the LDP map is presented as

$$H = (T_0, T_1, \dots, T_{n-1}). \quad (3)$$

To describe the LDP features, a depth silhouette image is divided into non-overlapping rectangle regions and the histogram is computed for each region. Furthermore, the whole LDP feature F is expressed as a concatenated sequence of histograms

$$F = (H^1, H^2, \dots, H^s) \quad (4)$$

where s represents the number of non-overlapped regions in the image.

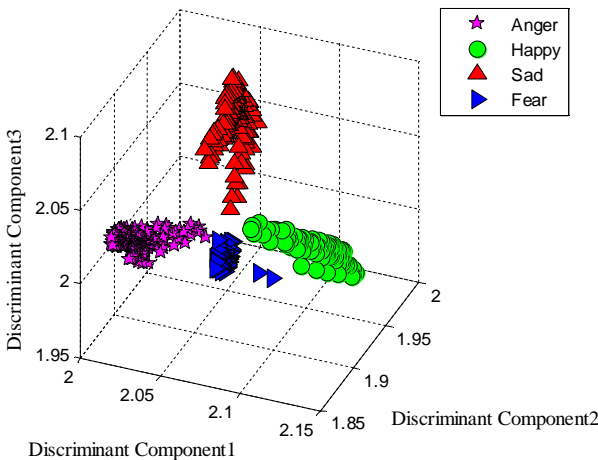


Fig. 7. 3-D plot of LDP-LDA features of depth faces from four expressions.

After analyzing the LDP descriptors of all the face depth images, it was noticed that there are some positions from all the positions corresponding to all the face images have values

greater than 0. Thus, it is better to consider only those positions of LDP descriptors of a face region and determine the standard dimension of the LDP descriptors for any face. However, the LDP features from the depth faces can be represented as D .

A. LDA on LDP Features

To obtain more robust features, LDA is performed on the LDP feature vectors F . Basically, LDA is based on class specific information which maximizes the ratio of the within, S_w and between, S_b scatter matrix. S_w and S_b that are computed by the following equations:

$$S_b = \sum_{i=1}^u V_i (\bar{\theta}_i - \bar{\theta})(\bar{\theta}_i - \bar{\theta})^T \quad (5)$$

$$S_w = \sum_{i=1}^u \sum_{m_k \in C_i} (\theta_k - \bar{\theta}_i)(\theta_k - \bar{\theta}_i)^T \quad (6)$$

where V_i represents the number of vectors in i^{th} class C_i and u the number of classes. $\bar{\theta}$ represents the mean of all vectors, $\bar{\theta}_i$ the mean of the class C_i and θ_k the vector of a specific class. The optimal discrimination matrix W_{LDA} is chosen from the maximization of ratio of the determinant of the between and within class scatter matrix as

$$W_{LDA} = \arg \max_{LDA} \frac{|W^T S_b W|}{|W^T S_w W|} \quad (7)$$

where W_D is the discriminant feature space. Thus, the LDP-LDA feature vectors of facial expression images can be obtained as follows.

$$J = F W_{LDA}^T \quad (8)$$

Fig. 7 shows an exemplar plot of 3-D LDA representation of the LDP features of all the facial expression depth images that shows a good separation among the representation of the depth faces of different classes.

B. Vector Quantization.

To analyze the sequential variations of the facial expression features, HMMs have been used in this work. As HMMs are usually trained and tested with discrete symbol sequences, the expression feature vectors are symbolized with the help of comparing with the codeword vectors of a codebook. To obtain a codebook, vector quantization algorithm is applied on the training feature vectors. In this work, the Linde, Buzo and Gray (LBG) algorithm has been utilized for codebook generation [21].

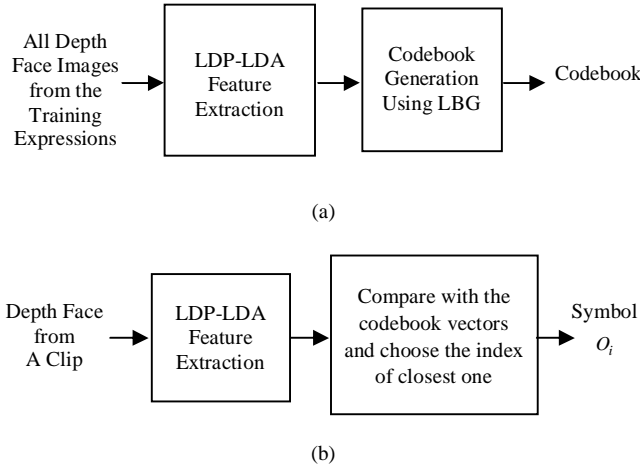


Fig.8. Step by step (a) codebook generation and (b) symbol selection.

When the codebook is obtained, the indices of the codewords are used as symbols to be used with discrete HMMs. As each face is converted to a symbol, an expression video clip of T face images will make T symbols after the vector quantization. Fig. 8 shows the steps for codebook generation and symbol selection process.

C. HMM for Expression Modeling and Recognition

To decode the depth information-based time-sequential facial expression features, discrete HMMs are employed. HMMs have been applied extensively to solve a large number of complex problems in various applications such as speech recognition [30].

An HMM is a collection of states where each state is characterized by transition and symbol observation probabilities. A basic HMM can be expressed as $H = \{Q, \pi, R, B\}$ where Q denotes possible states, π the initial probability of the states, R the transition probability matrix between hidden states where state transition probability r_{ij} represents the probability of changing state from i to j , and B observation symbols' probability from every state where the probability $b_j(O)$ indicates the probability of observing the symbols O from state j . If the number of activities is N then there will be a dictionary (H_1, H_2, \dots, H_N) of N trained models. We used the Baum-Welch algorithm for HMM parameter estimation as follows.

$$\xi_t(i, j) = \frac{\alpha_t(i) r_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^s \sum_{j=1}^s \alpha_t(i) r_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \quad (9)$$

$$\gamma_t(i) = \sum_{j=1}^s \xi_t(i, j) \quad (10)$$

$$\hat{r}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^T \gamma_t(i)} \quad (11)$$

$$\hat{b}_j(d) = \frac{\sum_{t=1}^{T-1} \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (12)$$

where $\xi_t(i, j)$ represents the probability of staying in a state i at time t and a state j at time $t+1$. $\gamma_t(i)$ is the probability of staying in the state i at time t . α and β are the forward and backward variables respectively that are calculated from transition and observation matrix. \hat{r}_{ij} is the estimated transition probability from the state i to the state j and $\hat{b}_j(d)$ the estimated observation probability of symbol d from the state j . s is the number of states used in the models. Further details regarding HMM can be obtained from [21]. Fig. 9 shows the structure and transition probabilities of a sad HMM before and after training.

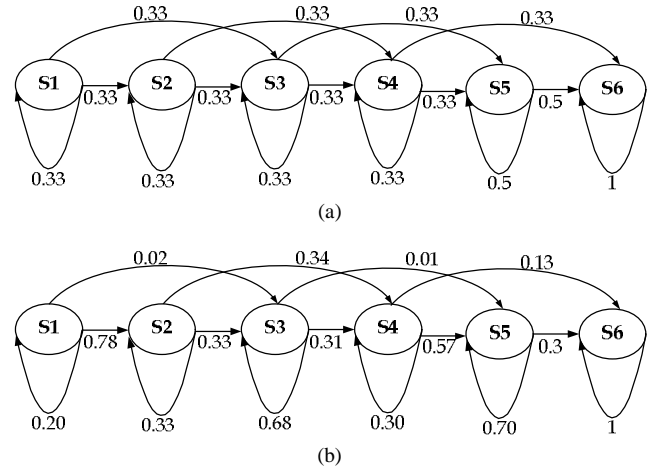


Fig. 9. A HMM transition probabilities for sad expression (a) before and (b) after training.

To test a facial expression video for recognition, the obtained observation sequence O from the corresponding depth image sequence is used to determine the proper model by highest likelihood L computation of all N trained expression HMMs as follows.

$$L = \arg \max_{k=1}^N (P(O | H_k)) \quad (13)$$

IV. EXPERIMENTS AND RESULTS

The FER database was built for six expressions: namely Surprise, Sad, Happy, Disgust, Anger, and Fear. Each expression video clip was of variable length and each expression in each video starts and ends with neutral expression. A total of 20 sequences from each expression were used to build the feature space. To train and test each facial expression model, 20 and 40 image sequences were applied respectively.

The average recognition rate using PCA on depth faces is 62.50% as shown in Table I. Then, we applied LDA on PCA features and obtained 65.83% average recognition rate as shown in Table II. As PCA-based global features showed poor recognition performance, we tried ICA-based local features for FER and obtained 83.33% average recognition rate as reported in Table III. To improve ICA features, we applied LDA on the ICA features and as shown in Table IV, the average recognition rate utilizing ICA representation on the depth facial expression images is 83.50%, which is higher than that of depth face-based FER applying PCA-based features. Then, LBP was tried on the same database that achieved the average recognition rate of 89.20% as shown in Table V. Furthermore, LDP was employed and achieved the better recognition rate than LBP i.e., 90.83% as shown in Table VI. Finally, LDP-LDA was applied with HMM and it showed its superiority over the other feature extraction methods achieving the highest recognition rate (i.e., 97.50%) as shown in Table VII. Fig. 10 shows the FER performances using different approaches on depth faces where proposed approach shows the superiority over others.

TABLE I
FER CONFUSION MATRIX USING DEPTH FACES WITH PCA.

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	50%	0	20	0	15	15
Happy	5	50	10	10	20	10
Sad	0	17.5	70	12.50	0	0
Surprise	0	0	0	80	15	5
Fear	0	5	25	10	60	0
Disgust	0	5	30	0	0	65
Average	62.50					

TABLE II
FER CONFUSION MATRIX USING DEPTH FACES WITH PCA-LDA.

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	50%	0	20	0	15	15
Happy	0	55	10	10	20	10
Sad	0	15	72.50	12.50	0	0
Surprise	0	0	0	85	10	5
Fear	0	5	25	5	65	0
Disgust	0	2.50	30	0	0	67.50
Average	65.83					

TABLE III
FER CONFUSION MATRIX USING DEPTH FACES WITH ICA.

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	80%	0	10	0	0	10
Happy	0	82.50	10	7.25	0	0
Sad	0	0	85	15	0	0
Surprise	0	0	0	85	15	0
Fear	0	5	15	0	85	0
Disgust	0	0	15	2.50	0	82.50
Average	83.33					

TABLE IV
FER CONFUSION MATRIX USING DEPTH FACES WITH ICA-LDA.

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	85%	0	5	0	0	10
Happy	0	85	10	5	0	0
Sad	0	0	85	15	0	0
Surprise	0	0	0	87.50	12.50	0
Fear	0	2.50	15	0	87.50	0
Disgust	0	0	15	0	0	85
Average	85.83					

TABLE V
FER CONFUSION MATRIX USING DEPTH FACES WITH LBP .

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	87.50%	0	12.50	0	0	0
Happy	10	90	0	0	0	0
Sad	0	0	90	0	10	0
Surprise	7.50	0	0	92.50	0	0
Fear	0	5	10	0	85	0
Disgust	0	0	10	0	0	90
Average	89.17					

TABLE VI
FER CONFUSION MATRIX USING DEPTH FACES WITH LDP.

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	90%	0	10	0	0	0
Happy	10	90	0	0	0	0
Sad	0	0	92.50	0	7.50	0
Surprise	7.50	0	0	92.50	0	0
Fear	0	5	7.50	0	87.50	0
Disgust	0	0	7.50	0	0	92.50
Average	90.83					

TABLE VII
FER CONFUSION MATRIX USING DEPTH FACES WITH LDP-LDA.

Expression	Anger	Happy	Sad	Surprise	Fear	Disgust
Anger	97.50%	0	2.50	0	0	0
Happy	5	95	0	0	0	0
Sad	0	0	97.50	2.50	0	0
Surprise	0	0	0	97.50	0	2.50
Fear	0	0	2.50	0	97.50	0
Disgust	0	0	0	0	0	100
Average	97.50					

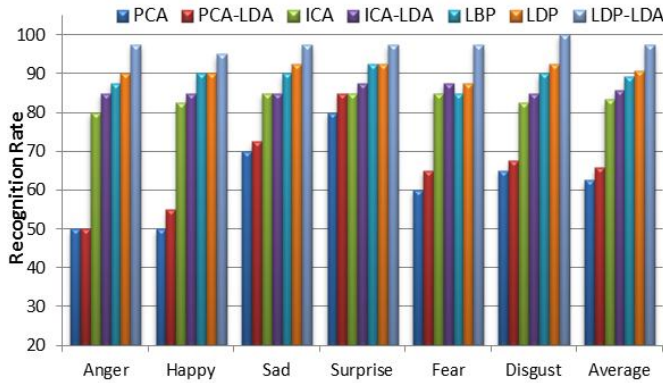


Fig. 10. FER performances using different approaches on depth faces.

V. CONCLUDING REMARKS

A depth video-based robust FER system has been proposed in this work using LDP-LDA features for facial expression feature extraction and HMM for recognition. The proposed method was compared with other traditional approaches and the recognition performance showed its superiority over others. However, the proposed system can be implemented in many systems such as smart home applications.

ACKNOWLEDGEMENT

This research was supported by the faculty research fund of Sungkyunkwan University, Republic of Korea.

REFERENCES

- [1] M. T. Rahman and N. Kertamavaz, "Real-Time Face-Priority Auto Focus for Digital and Cell-Phone Cameras," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1506–1513, 2008.
- [2] D.-S. Kim, I.-J. Jeon, S.-Y. Lee, P.-K. Rhee, and D.-J. Chung, "Embedded Face Recognition based on Fast Genetic Algorithm for Intelligent Digital Photography," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 726–734, 2006.
- [3] C. Padgett and G. Cottrell, "Representation face images for emotion classification," *Advances in Neural Information Processing Systems*, vol. 9, Cambridge, MA, MIT Press, 1997.
- [4] S. Mitra and T. Acharya, "Gesture Recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.
- [5] G. Donato, M. S. Bartlett, J. C. Hagar, P. Ekman, and T. J. Sejnowski, "Classifying Facial Actions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.
- [6] P. Ekman and W. V. Priesen, "Facial Action Coding System: A technique for the Measurement of Facial Movement," *Consulting Psychologists Press*, Palo Alto, CA, 1978.
- [7] M. Meulders, P. D. Boeck, I. V. Mechelen, and A. Gelman, "Probabilistic feature analysis of facial perception of emotions," *Applied Statistics*, vol. 54, no. 4, pp. 781–793, 2005.
- [8] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vision Research*, vol. 41, pp. 1179–1208, 2001.
- [9] S. Dubuisson, F. Davoine, and M. Masson, "A solution for facial expression representation and recognition," *Signal Processing: Image Communication*, vol. 17, pp. 657–673, 2002.
- [10] I. Buciu, C. Kotropoulos, and I. Pitas, "ICA and Gabor Representation for Facial Expression Recognition," in *Proceedings of the IEEE*, pp. 855–858, 2003.
- [11] F. chen and K. Kotani, "Facial Expression Recognition by Supervised Independent Component Analysis Using MAP Estimation," *IEICE Transactions on Information and Systems*, vol. E91-D, no. 2, pp. 341–350, 2008.
- [12] A. Hyvarinen, J. Karhunen, and E. Oja, "Independent Component Analysis," *John Wiley & Sons*, 2001.
- [13] Y. Karklin and M.S Lewicki, "Learning higher-order structures in natural images," *Network: Computation in Neural Systems*, vol. 14, pp. 483–499, 2003.
- [14] M. S. Bartlett, G. Donato, J. R. Movellan, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Face Image Analysis for Expression Measurement and Detection of Deceit," in *Proceedings of the Sixth Joint Symposium on Neural Computation*, pp. 8–15, 1999.
- [15] C. Chao-Fa and F. Y. Shin, "Recognizing Facial Action Units Using Independent Component Analysis and Support Vector Machine," *Pattern Recognition*, vol. 39, pp. 1795–1798, 2006.
- [16] A. J. Calder, A. W. Young, and J. Keane, "Configural information in facial expression perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 2, pp. 527–551, 2000.
- [17] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp.200–205, 1998.
- [18] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face Recognition by Independent Component Analysis," *IEEE Transaction on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002.
- [19] C. Liu, "Enhanced independent component analysis and its application to content based face image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 34, no. 2, pp. 1117–1127, 2004.
- [20] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, pp. 295–306, 1998.
- [21] M. Z. Uddin, J. J. Lee, and T.-S. Kim, "An Enhanced Independent Component-Based Human Facial Expression Recognition from Video," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2216–2224, 2009.
- [22] T. Ojala, M. Pietikäinen, T. Mäenpää, "Multiresolution gray scale and rotation invariant texture analysis with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 971–987, 2002.
- [23] C. Shan, S. Gong, P. McOwan, "Robust facial expression recognition using local binary patterns," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 370–373, 2005.
- [24] C. Shan, S. Gong, and P. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, pp. 803–816, 2009.
- [25] T. Jabid, M. H. Kabir, O. Chae, "Local Directional Pattern (LDP) A Robust Image Descriptor for Object Recognition", in *Proceedings of the IEEE Advanced Video and Signal Based Surveillance (AVSS)*, pp. 482–487, 2010.
- [26] Y. Zhu, L. C. De Silva, and C. C. Ko, "Using moment invariants and HMM in facial expression recognition," *Pattern Recognition Letters*, vol. 23, no. (1-3), pp. 83–91, 2002.
- [27] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, pp. 160–187, 2003.
- [28] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream HMMs," *IEEE Transaction on Information and Security*, vol. 1, pp. 3–11, 2006.
- [29] G. J. Iddan and G. Yahav, "3D imaging in the studio (and elsewhere...)," in *Proceedings of SPIE*, vol. 4298, pp 48–55, 2001.
- [30] L. R. Rabiner, "A Tutorial on Hidden Markov Modes and selected application in speech recognition," in *Proceedings of IEEE*, vol. 77, pp. 257–286, 1989.

Performance Analysis of Face Detection Algorithms for Efficient Comparison of Prediction Time and Accuracy

Seunghui Cha, Jong Wook Kwak, and Wookhyun Kim
Department of Computer Engineering
Yeungnam University, Gyeongsan, 712-749, Republic of Korea

Abstract - *Face detection is one of challenges in image processing. It is necessary to compare two or more face detection algorithm to effectively select candidate algorithms based on their detection time and accuracy. In this paper we analyze three face detection algorithms and then provide accuracy and performance of each algorithm. Candidate algorithms for face detection method are skin color, haar feature and facial feature. Based on our analysis, we have checked that each algorithm has their unique characteristics and our experimental results show that, depend on algorithms, their detection accuracy varies 66%, 87%, 93%, respectively.*

Keywords: face detection, face extraction, skin color detection, haar feature, adaboost, facial feature, classifier evaluation.

1 Introduction

Face detection is recognized as an integral part in the field of computer vision and image processing. The application of computer vision is robot, medical image analysis, human-computer interaction, scene reconstruction, video tracking, object recognition, motion estimation and so on. In particular, extraction and recognition of human face have received special attention and have developed those techniques rapidly. The face detection in image processing is important because detection is an important first step in recognition. In this paper we analyze three algorithms of face detection in terms of their efficiency. Our purpose is to examine the advantages and disadvantages of each algorithm with the 1000 image dataset [1]. The result of experiment will be discussed based on speed and accuracy of face detection. The characteristic of the image data used in the experiment has a variety of human images on a complex background and various illuminations. So the images have the more complex environment than normal images in detecting face. In detail, three algorithms are as follows. The first algorithm detects a face by using skin color [2]. Skin color is effective for face detection and it is invariant in geometric variations [5]. The second algorithm detects a face by using the Haar feature classifier of adaboost [3]. Haar feature increases the speed and accuracy on face detection greatly. The third algorithm detects a face by using a facial feature [4]. Facial feature is an important part in the field of face recognition. Based on these

face detection algorithms, we introduce main characteristics and differences for each one. The rest of this paper is structured as follows. After discussing related work in section 2, section 3 describes face detection using skin color, Haar feature classifier, a hierarchical face identification system using facial components. Section 4 shows the experimental results. In section 5, the conclusion is discussed and future works are presented.

2 Related works

During the past decade, many methods for face detection have been studied. Many techniques for face detection are classified as follows. The main categories of detection method are knowledge based method [6][7], image based method, feature based method, template matching method [8] and so on. Knowledge based method deals with feature of face and relationship of face. Image based method deals with whether the face pattern is face or non-face. Feature based method is used to find feasible matches between object features and image features. Template matching method deals with comparison of image in standard image set. Other classification method is as follows. There are adaboost classifier, Linear Discriminant Analysis (LDA) [9], Principal Component Analysis (PCA), Independent Component Analysis (ICA) [10], Kernel Principal Component Analysis (KPCA) [11], Support Vector Machine (SVM) [12] and so on. Viola and Jones proposed the haar feature based boosted detector (adaboost). This method has a higher extraction rate and a fast speed on image processing. The learning algorithm designed using an adaboost classifier is simple and effective. Feature extracts using vector by Linear Discriminant Analysis (LDA) was also proposed. This method calculates the most accurate vector between two classes. It seeks vectors among classes in the best determined space and captures global geometrical structure information of the data. Many LDA extensions have been developed [13]. Principal Component Analysis (PCA) finds vector space of maximum variance. This method calculates vectors with the best associated variance. It finds the sub space with the basic vectors to consider the direction of the largest variance in the original space. Independent Component Analysis (ICA) executes with linear mixing of signals. This method is calculated for the separation of the multivariate signal. It is the linear combination of independent source on the part of the real

evaluated number. In addition, Kernel Principal Component Analysis (KPCA) is an extension of Principal Component Analysis (PCA). This method is a non-linear extension of PCA and input space is made into the feature space of a non-linear space. Support Vector Machine (SVM) is used for classification and regression analysis. This method is derived from the statistical learning theory and the function of the input is used for high dimensional feature space defined by the kernel function. The types of features have geometric feature data and appearance feature data in facial expression recognition [14]. There are shape and position of the feature in geometric features and there are the wrinkles, bulges and furrows in appearance features. There are micro-patterns of the facial expressions in appearance features. Geometric features provide much information and are sensitive to noise in facial expression recognition.

3 Algorithms of face detection

In this section, characteristics of three face detection algorithm are introduced.

3.1 Face detection using skin color

Detection of the face is used with skin color classifier. This method detects face quickly. Skin color classifier perform that skin color distinguishes from background. A variety of color space presents information in image processing field. Mostly, four kinds of color space are often used.

3.1.1 Color space

Typical types of color space are as follows [15].

a) RGB: Colors are classified into three primary colors. Colors are Red (R), Green (G) and Blue (B). RGB is a device-dependent color model.

b) HSV: Color perception has characteristics of three. Colors are Hue (H), Saturation (S) and intensity Values (V). There is the nonlinear transformation between RGB and HSV. her similar color space is HIS, HLS and HCl. HSL and HSV have the most common cylindrical-coordinate representations of points in RGB color model.

c) YCbCr: The color is classified by the luminance (Y channel) and chrominance (Cb and Cr channels). There is the linear transformation between RGB and HSV. Other similar color space is YIQ and YUV. It is a way of encoding RGB information.

d) CIE-Lab: It is designed to approximate perceptually uniform color spaces (UCSs). CIE-Lab color space is related to RGB color space through a highly nonlinear transformation. Examples of similar color spaces are CIE-Luv and Farnsworth UCS.

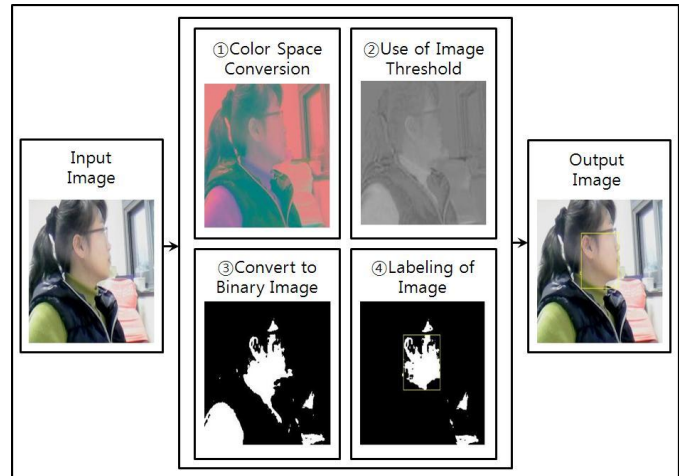


Fig. 1. Face Detection by Skin Color

3.1.2 Color space conversion

CIE equalizes to quantify brightness between the device-independent color spaces. Color space of $L^*a^*b^*$ is generated by the CIE and is normalized to a reference white point. L^* represents the lightness, a^* represents the red and green value, and b^* represents the yellow and blue value of the color, respectively [16]. The color quantifies the visual differences by converting of color space.

3.1.3 Image threshold

Image is used to Otsu's method. It method performs clustering-based image threshold [17][18]. The threshold value is shown in equation (1). T value is grey level value. The image of gray scale is changed to binary image [19]. The pixel color is changed based on whether pixel value is greater than the T value or pixel value is less than the T value.

$$A \text{ pixel becomes } \begin{cases} \text{white if its gray level is } > T \\ \text{black if its gray level is } \leq T \end{cases} \quad (1)$$

The threshold value is shown in equation (1). n_i in equation (2) is number of pixels. N is total number of pixel. P_i is with value of i level.

$$P_i = \frac{n_i}{N} \quad (2)$$

In equation (3), L is number of gray scale. $L-1$ is maximum value of gray scale.

$$\omega(\kappa) = \sum_{i=0}^{\kappa} P_i, \quad \mu(\kappa) = \sum_{i=\kappa+1}^{L-1} P_i \quad (3)$$

In equation (4), the difference between $w(k)$ and $u(k)$ is maximized.

$$\omega(\kappa) + \mu(\kappa) = \sum_{i=0}^{L-1} P_i = 1 \quad (4)$$

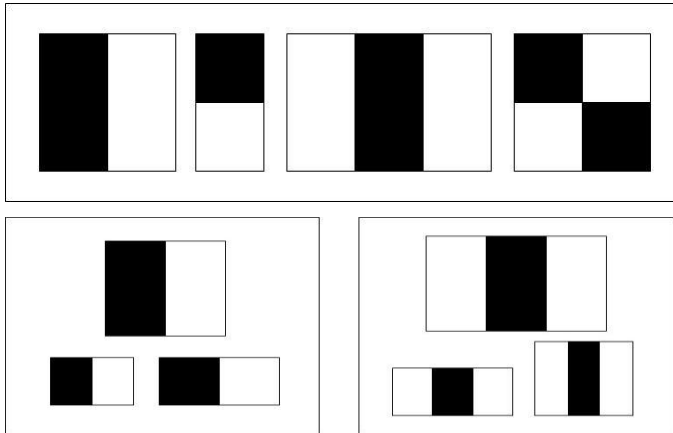


Fig. 2. Haar Feature

Equation (5) is defined as the average of the image.

$$\mu T = \sum_{t=0}^{L-1} I P_t \quad (5)$$

Equation (6) searches the maximum of k value.

$$\frac{(\mu T \omega(\kappa) - \mu(\kappa))^2}{\omega(\kappa) - \mu(\kappa)} \quad (6)$$

3.1.4 Connected-component labeling

Labeling of connected components is based upon the heuristic. Labeling is used to detect the connected region [20].

- a) Label assigns and label records in a local table.
- b) Label assigns and label records in equivalence classes.

The process of a) and b) is repeated. Fig. 1 shows the picture of face detection by using skin color. First, the color space converts to L*a*b* color space with input image and then the image is converted to binary image with threshold value. Finally image labeling is used to detect a face.

3.2 Face detection using haar feature classifier

Haar is an image representation called an integral image that allows for very fast feature evaluation. It is a method for constructing a classifier by selecting a small number of important features using adaboost [21].

3.2.1 Features

Feature is used instead of pixel because feature-based system is faster than pixel-based system.

- a) Simple feature: Adaboost algorithm is able to select a good feature in given features and uses features of three types (two-rectangle, three-rectangle, four-rectangle).

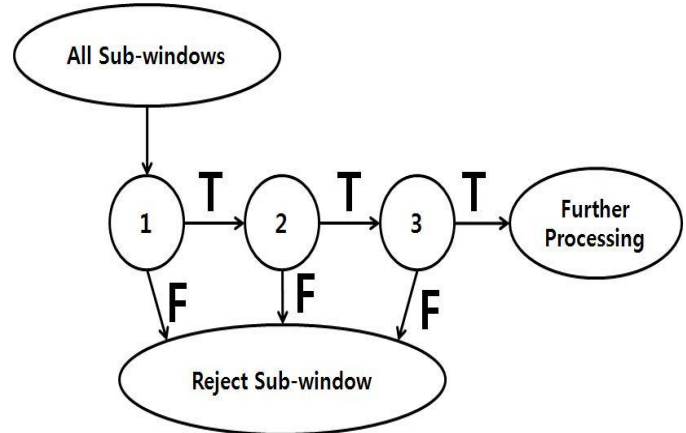


Fig. 3. Schematic Depiction of the Detection Cascade

Simple feature is usually called haar feature. The haar feature has rectangle feature as shown in Fig. 2.

- b) Haar feature: The value of two-rectangle feature is the difference between the sum of the pixels within two rectangular regions. Three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally four-rectangle feature computes the difference between diagonal pairs of rectangles.

3.2.2 Classifier

This algorithm learns classification function.

- a) Weak classifier: Features are combined to make an efficient classifier. The weak classifier formula (7) is expressed as a functional form about classifier. H is weak classifier. X is input of sub window. F has a Haar feature. P has value of parity. θ has value of threshold.

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{if } pf(x) < p\theta \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

- b) Strong classifier: Strong classifier is created by combining the collection of weak classification functions.

3.2.3 Cascade algorithm

It constructs a cascade of classifiers which achieves increased detection performance.

- a) Composition: To make a fast, reliable and efficient face detector, a cascaded classifier is constituted. A classifier is configured as a single cascade in order to reduce of computational time and increase of detection performance.
- b) Boosted classifier: This is configured as a boosted classifier which detects positive instances and rejects negative sub-windows.
- c) Threshold: This obtains a effective face filter that the strong classifier adjusts a threshold value to minimize the

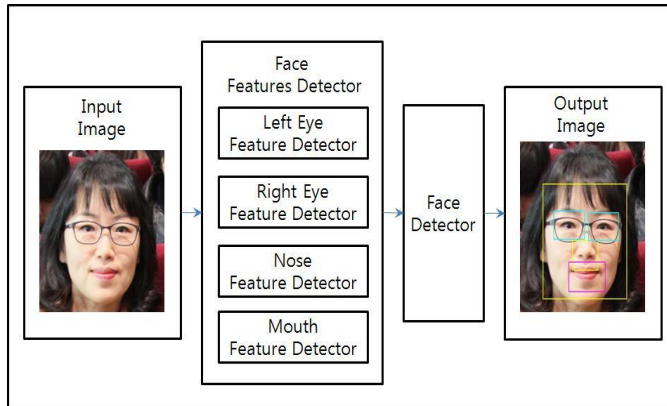


Fig. 4. Face Detection by a Facial Feature

false negative. Fig. 3 shows schematic depiction of the detection cascade.

3.3 Face detection using facial feature

This method detects important features of face [22]. Important features are left eye, right eye mouth and nose. After locating features, a face is detected. In order to detect the facial feature, the following process is performed.

a) Convert to binary image: First, an image is converted to a binary image. This is the same as the skin color method.

b) Labeling of image: Labeling operations are performed on the image. This process is also the same as face detection using skin color.

c) Haar feature and cascade: Facial feature is based on haar feature and cascade. This is the same as haar feature face detection.

d) Detection of facial components: Features of eyes, nose, and mouth are extracted. The diagram of a face detection using facial feature is shown in Fig. 4. Equations to extract the facial features are as follows [23]. Equation (8) shows the m th feature of the q th face image ($q=1, \dots, Q$) of the k th ($k = 1, \dots, K$).

$$\bar{f}_m(k) = \frac{1}{Q} \sum_{q=1}^Q f_m(k, q) \quad (8)$$

TABLE 1 Detection Rate and Average Detection Time

Face Detector	Detection Rate	Average Detection Time
Skin Color	66 %	0.410 sec
Haar Feature	87 %	0.443 sec
Facial Feature	93 %	1.669 sec

Equation (9) shows the average of features over the database.

$$\bar{f}_m = \frac{1}{KQ} \sum_{k=1}^K \sum_{q=1}^Q f_m(k, q) \quad (9)$$

Equation (10) shows the ratio of m th feature point in the feature vector.

$$r_m = \frac{\sum_{k=1}^K (\bar{f}_m(k) - \bar{f}_m)^2}{\sum_{k=1}^K \sum_{q=1}^Q (f_m(k, q) - \bar{f}_m(k))^2} \quad (10)$$

Consequently, the ranking features are obtained with r_m and $f_{r(m)}$ in the m th ranking feature. Ranking of M features are $\{f_{r(1)}, f_{r(2)}, \dots, f_{r(M-1)}, f_{r(M)}\}$.

4 Experiment

In section, we present the result of experiment and analysis of three face detection algorithms with 1000 image dataset. We use face detection algorithms using skin color, haar feature and facial feature.

4.1 Test on dataset

We chose the image dataset with complex background among thousands of images because all face detectors extract well on images with simple background. So we experiment on the image dataset with complex background image. We use images selected from Fddb (Face Detection Data Set and Benchmark Home, <http://vis-www.cs.umass.edu/fddb/>) [1]. Each image has different characteristics and various scenes as follows. There are front of face and side of face in face image and there are many people or one person. There are person dressed in a short sleeve, a long sleeve and a sleeveless shirt. And there are face hidden by an object, reflective glasses, background color similar to the face and blurred face in image.

4.2 Experimental result

TABLE 2 Comparison of Three Algorithms

Feature	Skin Color	Haar Feature	Facial Feature
Many People	×	△	○
All Skin Color	×	×	○
Side Face	△	△	○
Sleeveless Shirt	△	△	○
Face hidden by an Object	△	○	△
Reflective Glasses	×	○	×
Blurred Image	○	△	○
Similar Color	△	○	○

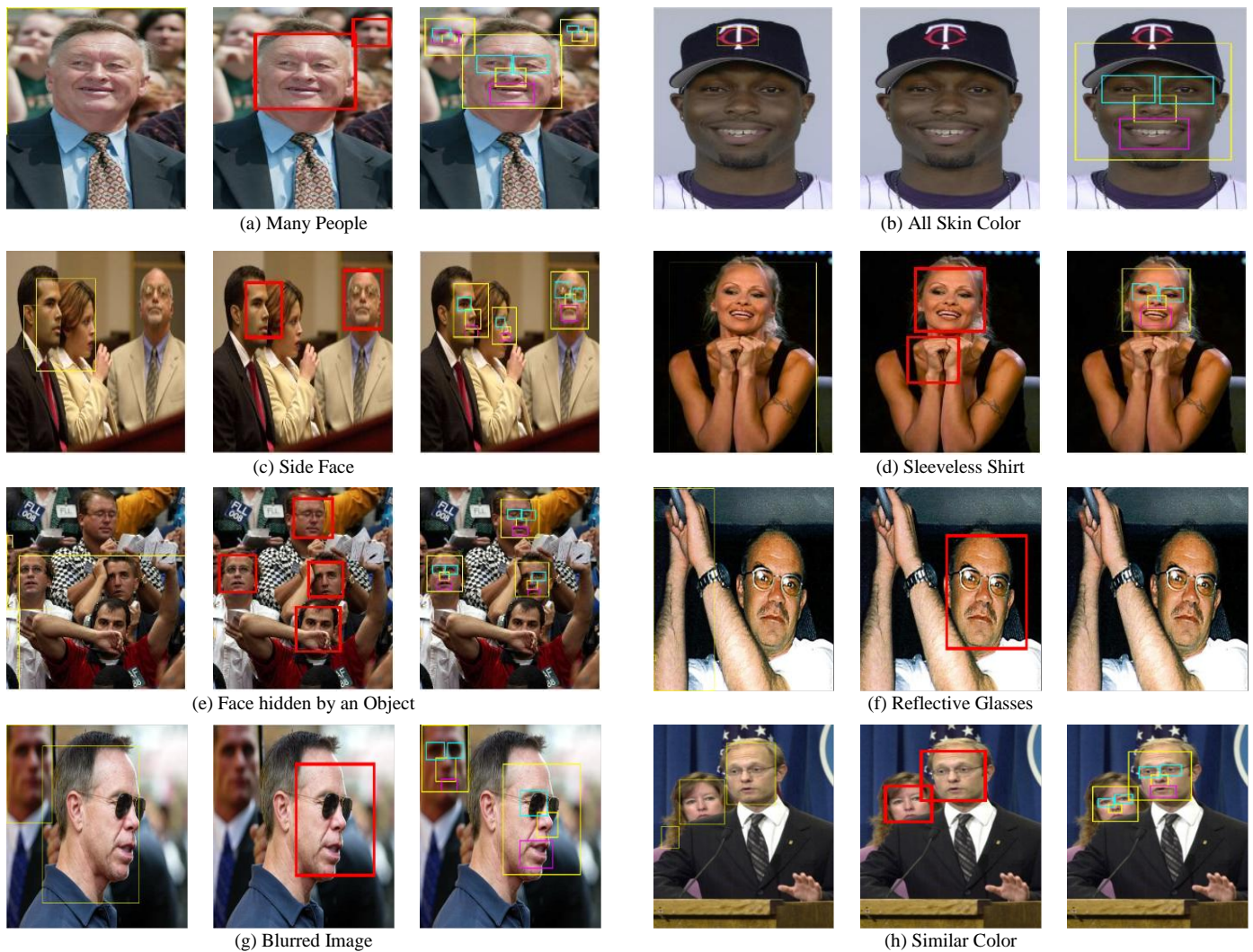


Fig. 5. Face Detection using three Algorithms

4.2.1 Comparison of detection rate and detection time

The face detection rates and the average detection times of each algorithm are shown in Table 1. The detection rate by using skin color is 66% and the average detection time by using skin color is 0.410sec. The detection rate by using haar feature is 87% and the average detection time by using haar feature is 0.443sec. The detection rate by using facial feature is 93% and the average detection time by using facial feature is 1.669sec. The facial feature detector has the highest extraction rate but has the slowest extraction time. That is, the accuracy of face detection is high but time efficiency is low. In this paper, face detection rate was defined as follows. If the face is detected, the value is 1. If the face is not detected, the value is 0. If face and others are detected at the same time, the value is also 0.

4.2.2 Comparison of three algorithms

Fig. 5(a) shows images which are many people in the background. Fig. 5(b) is an image with a dark skin color. Fig. 5(c) is a side view of face. Fig. 5(d) shows an image of a

person dressed in a sleeveless shirt. Fig. 5(e) shows face hidden by an object. Fig. 5(f) is an image that is reflected the glasses. Fig. 5(g) shows an image of the blurred face. Fig. 5(h) shows background color similar to the face. The results of the face detection are shown in Table 2. Characteristic of symbols \circ , Δ , \times in Table 2 are as follows. The symbol \circ indicates that the face detection is perfectly well. The symbol Δ indicates that it is detected face well but non-face is also detected together. The symbol \times indicates that face is not detected.

a) Using skin color: The face detection by using skin color has disadvantage of low accuracy.

b) Using haar feature: The face detection by using haar feature shows the strength in the fig. 5(e) with hidden face by an object, fig. 5(f) with reflected glasses and fig. 5(h) with background color similar to the face. However, this method shows a weakness when the skin color is composed a variety of colors.

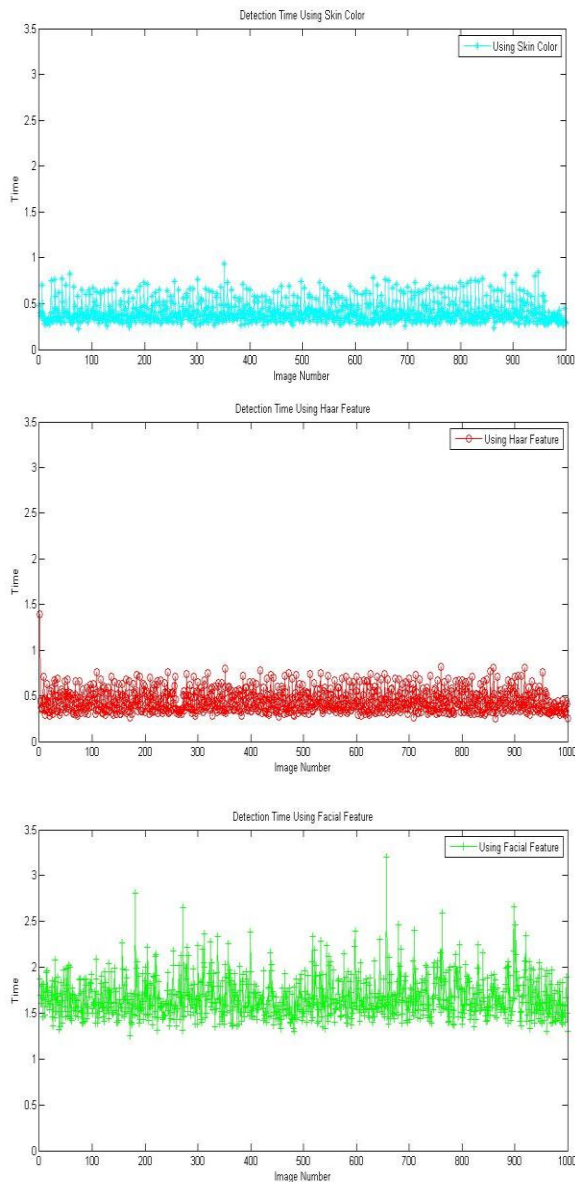


Fig. 6. Face Detection Time using Three Algorithms

c) Using facial feature: The face detection by using facial feature has the strength as following conditions. Fig. 5(a) with many people, fig. 5(b) with a variety of skin colors, fig. 5(c) with face of side view, fig. 5(d) with person dressed a sleeveless shirt, fig. 5(g) with a blurred image, fig. 5(h) with background color similar to skin color. These conditions have shown the strength. The method has the strength generally. However, this method has a weakness as following conditions. The image in fig. 5(e) has hidden face by an object and fig. 5(f) has reflected glasses. These conditions have shown the weakness.

Our experimental result indicates that first and second methods have fast time and low accuracy. Third method has slow time but high accuracy. Particularly third method has strength on face detection under complex background. Finally,

fig. 6 shows comparison of detection time using three algorithms. First figure is detection time using skin color, second figure shows detection time using haar feature. Third figure shows detection time using facial feature. As shown in the chart, third figure has the most time on face detection. Detection time of first figure is similar to that of second figure.

5 Conclusion

Face detection is one of challenges in image processing. It is necessary to compare two or more face detection algorithm to effectively select candidate algorithms based on their detection time and accuracy. In this paper, we present experimental results of face detection algorithms. The analysis of each algorithm is provided to compare the accuracy and performance on image dataset. Our comparative study approach addresses the accuracy and the performance of algorithms to integrate more detailed information on complex images. We have demonstrated the scalability of our experiment by applying it on 1000 image sets with three algorithms. Our experiments are shown by using skin color, haar feature, facial feature for face detection. Based on our analysis, we have checked that each algorithm has their unique characteristics. The face detection using facial feature has the slowest detection time but it has high accuracy. The face detection by using facial feature shows a robust detection rate in a variety of conditions. But the method using facial feature shows a weak detection when a face overlaps with other things. The method using skin color has the fastest detection time but the method has the low accuracy. In summary, our experimental results show that, depend on algorithms, their average prediction time varies and their average accuracy also varies 66%, 87%, 93%, respectively. In the future, we plan to add an option in the existing algorithms to improve the performance.

6 References

- [1] Vedit Jain and Erik Learned-Miller, "FDDB: A Benchmark for Face Detection in Unconstrained Settings," University of Massachusetts Amherst, Technical Report, Feb. 2010. <http://vis-www.cs.umass.edu/fddb/>.
- [2] Christophe Garcia and Georgios Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis," *IEEE Trans. MULTIMEDIA*, vol. 1, no. 3, Sep. 1999.
- [3] Paul Viola, Michael J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, Issue 2, pp. 137-154. May 2004.
- [4] Mehrtash T Harandi, Malid Nili Ahmadabadi, and Babak N. Araabi, "A Hierarchical Face Identification System Based on Facial Components," *Proc. IEEE Int'l Conf. Computer System and Applications*, pp. 669-675, May 2007.

- [5] Erdem, C.E., Ulukaya, S., Karaali, A. and Erdem, A.T. "Combining Haar Feature and skin color based classifiers for face detection," Proc. IEEE Int'l Conf. Speech and Signal Processing, pp. 1497-1500, May 2011.
- [6] Thai Hoang Le "Applying Artificial Neural Networks for Face Recognition," Advances in Artificial Neural Systems, vol. 2011, Jan. 2011.
- [7] MacGregor, Robert "Using a description classifier to enhance knowledge representation," IEEE Expert vol. 6, pp. 41-46, June 1991.
- [8] Tarun Kumar, Kushal Veer Singh and Shekhar Malik "Artificial neural network in face detection," International Journal of Computer Applications, vol. 14, no. 3, Jan. 2011.
- [9] Martinez, Aleix M. and Kak, A.C. "PCA versus LDA," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, pp. 228-233, Feb. 2001.
- [10] Huy Nguyen and Rong Zheng, "Binary Independent Component Analysis With or Mixture," IEEE Trans. Signal Processing, vol. 59, pp. 3168-3181, July 2011.
- [11] Kwang In Kim, "Face recognition using kernel principal component analysis," Proc. IEEE Conf. Neural Networks for Signal Processing, vol. 9, pp. 337-343, Feb. 2002
- [12] Karatzoglou and Alexandros, "Support Vector Machines in R," Journal of Statistical Software, vol. 15, pp. 1-28, April 2006.
- [13] Di Zhang, Jiazhong He, Yun Zhao, Zhongliang Luo, Minghui Du "Global plus local: A complete framework for feature extraction and recognition," Pattern Recognition, vol. 47, pp.1433-1442, March 2013.
- [14] Anima Majumder, Laxmidhar Behera and Venkatesh K. Subramanian "Emotion recognition from geometric facial features using self-organizing map," Pattern Recognition, vol. 47, Issue 3, pp. 1282-1293, March 2014.
- [15] Phung, S.L., Bouzerdoum, A. and Chai, D., Sr. "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison," IEEE Trans. Pattern Analysis and machine Intelligence, vol. 27, pp. 148-154, Jan. 2005.
- [16] Chen-Chung Liu and Guan-Nan Hu, "A re-coloring algorithm for a color image using statistic scheme in CIE $L^*a^*b^*$ color space," Proc. IEEE Int'l Conf. Computer Communication Control and Automation, vol. 1 pp. 240-243, May 2010.
- [17] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," IEEE Trans. Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66, 1979.
- [18] M. Sezgin and B. Sankur "Survey over image thresholding techniques and quantitative performance evaluation," Journal of Electronic Imaging, vol. 13, pp. 146-165, Jan. 2004.
- [19] Alasdair McAndrew "Introduction to digital image processing with matlab," 2004.
- [20] H. Samet and M. Tamminen "Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintreees". IEEE Trans. Pattern Analysis and Machine Intelligence," vol. 10, pp. 579-586, Jul. 1998.
- [21] Zhengming Li, Lijie Xue and Fei Tan "Face Detection in Complex Background Based on Skin Color Features and Improved AdaBoost Algorithms," Proc. IEEE Int'l Conf. Progress in Informatics and Computing, vol. 2, pp. 723-727, Dec. 2010.
- [22] M. Castrillón, O. Déniz, C. Guerra and M. Hernández "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," Journal of Visual Communication and Image Representation, vol.18, pp.130-140, April 2007.
- [23] Valuvanathorn, S., Nitsuwat, S. and Mao Lin Huang "Multi-feature face recognition based on PSO-SVM," Proc. IEEE Int'l Conf. ICT and Knowledge Engineering, pp. 140-145, Nov. 2012.

A Novel SVM based Pedestrian Detection Algorithm via Locality Sensitive Histograms

Zhihui Wang¹, Sook Yoon², Changpyo Hong¹, and Dong Sun Park³

¹ Division of Electronic and Information Engineering, Chonbuk National University, Jeonju, South Korea

² Department of Multimedia, Mokpo National University, Jeonnam, South Korea

³ IT Convergence Research Center, Chonbuk National University, Jeonju, South Korea

Abstract - This paper present a novel support vector machine (SVM) based pedestrian detection algorithm with locality sensitive histograms. Locality sensitive histograms are rather effective and robust to represent different categories of objects, which can be calculated in time linear in the image size and the number of bins. The proposed pedestrian detection algorithm combines SVM and locality sensitive histograms, and this algorithm is powerful and robust to pedestrians with different scales. We show in our experiment that locality sensitive histograms perform better or comparable with other state-of-the-art features.

Keywords: Pedestrian detection; support vector machine; locality sensitive histograms; feature extraction

1 Introduction

Pedestrian detection is a very challenging task in computer vision. Numerous pedestrian detection algorithms have been proposed during the past decades [1-5]. However, the detecting efficiency and accuracy has yet to be improved. Generally speaking, the pedestrian detection is influenced by many internal and external factors. The internal factors contain pedestrian posture and appearance changes, and dressing styles. The external factors contains illumination and background variation, heavy occlusions. All these factors are difficult to deal with in real applications.

For these existing object detection systems, machine learning algorithms have been widely used, such as support vector machine (SVM) [6, 7], cascade AdaBoost algorithm (CAB) [8, 9], and random forest algorithm (RF) [10, 11]. Among them, the SVM has always been utilized then others. However, a proper feature vector should be extracted to represent pedestrians and backgrounds, which is regarded as the input of these machine learning algorithms. These features are crucial important to enhance the performances of these machine learning algorithms to distinguish pedestrians against other surrounding backgrounds. Therefore, how to combine these machine learning algorithms and corresponding features are rather important during the procedure of pedestrian detection.

There are so many efficient and robust features have been proposed for various purposes, which are rather powerful in corresponding applications and have been widespread concerned. Random Haar-like (RHL) feature [12] is convenient to extract, which can be regarded as one sparse representation method, and this feature is mainly employed in the field of visual tracking. Local binary patterns (LBP) [13] and histogram of oriented gradients (HOG) [5] are two widely used image features, however, both of them have limitations in their specific applications. Concerning this issue, feature selection is rather important to represent samples in different classes and directly affect the final detecting accuracy.

A novel image feature called locality sensitive histograms (LSH) [14] has been proved suitable to distinguish the target object in visual tracking algorithms [15-18]. The LSH feature takes into account contributions from every pixel adaptively and sufficiently. The efficient illumination invariant feature extracted based on LSH feature is rather robust to illumination changes. Considering that locality sensitive histograms can be extracted so fast by the strategy of integral image and have high performance in the field of pattern recognition, in this paper, we employ LSH feature to represent pedestrians and corresponding negative samples. Combined with LSH feature, SVM algorithm is chosen to classify these pedestrian and negative samples. The performance of this combination has been proved robust and accurate in the field of pedestrian detection with different image scales in our experiments.

The rest of this paper is organized as follows. We start by introducing the structure of the proposed system in Section 2, and the experimental comparison of the proposed pedestrian detection system via LSH feature with other state-of-the-art features is demonstrated in Section 3. Finally, we conclude the proposed system and discuss the superiority of LSH feature over other features in Section 4.

2 Proposed Pedestrian Detection System

Feature extraction and classification model selection are two of these essential steps of object classification and detection in computer vision. In this section, we discuss the LSH feature extraction and SVM algorithm in details

separately, which are employed in this proposed detection system.

2.1 Locality sensitive histograms (LSH)

Locality sensitive histograms [14] is a novel feature proposed for tracking by detection algorithm, which takes into account contributions from every pixel in a image instead of from pixel inside local neighborhoods as the local histograms. This feature is calculated in the strategy of integral image, therefore, the efficiency of LSH feature extraction can be guaranteed. Locality sensitive histograms are rather convenience to extract with computational complexity $O(NB)$, where N is the number of pixels and B is the number of bins. Let $H_p^E(b)$ represent the b th bin of LSH feature estimated at the pixel p of given image I :

$$H_p^E(b) = \sum_{q=1}^W \alpha^{|p-q|} Q(I_q, b), \quad b=1, \dots, B, \quad (1)$$

where W is the number of pixels, $Q(I_q, b)$ is zero expect when intensity value I_q of pixel q belongs to bin b , and $\alpha \in (0, 1)$ is the decreasing weight based on the distance between p and q .

Similar to the integral histogram, the LSH feature can be calculated efficiently. Take a simple example of 1D image, $H_p^E(b)$ can be calculated as:

$$H_p^E(b) = H_p^{E, \text{left}}(b) + H_p^{E, \text{right}}(b) - Q(I_p, b), \quad (2)$$

where

$$H_p^{E, \text{left}}(b) = Q(I_p, b) + \alpha H_{p-1}^{E, \text{left}}(b), \quad (3)$$

$$H_p^{E, \text{right}}(b) = Q(I_p, b) + \alpha H_{p+1}^{E, \text{right}}(b), \quad (4)$$

Therefore, the LSH feature can be estimated in time linear in the image size and the number of bins without any repetitive computation. In order to deal with illumination variation, illumination invariant features are calculated based on image transform and LSH feature, which has been deduced in details in [14].

2.2 Support vector machine (SVM)

Support vector machine is a rather powerful machine learning algorithm, which has been utilized widely for classification and regression analysis problems. Based on structural risk minimization theory, the performance of SVM classifier is accurate and robust enough to distinguish target objects from various of candidates. Here we focus on the binary classification problems of pedestrian detection. During this issue, SVM classifier can find a hyperplane segment the pedestrian and non-pedestrian samples with largest margin between them.

Let $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in R^d$ and $y_i \in \{\pm 1\}$ be training dataset, where x_i is the input vector, and -1 is non-pedestrian label and

1 is the pedestrian label. The decision boundary should classifier all these training samples with highest classification accuracy. The classification hyperplane can be expressed as $\omega \cdot x + b = 0$. Thus, the optimal hyperplane with minimum distance to the origin can be expressed as:

$$\begin{aligned} \arg \min_{(\omega, b)} \quad & \phi(\omega) = \frac{1}{2} \|\omega\|^2 \\ \text{subject to} \quad & y_i [\omega \cdot x_i + b] \geq 1, \quad i=1, \dots, n \end{aligned} \quad (5)$$

The above constrained optimization problem can be transformed to be dual form using Karush–Kuhn–Tucker condition and Lagrange optimizing method. Therefore, these samples can be classified using the following discrimination function:

$$f(x) = \text{sgn} \left\{ \omega \cdot x + b \right\} = \text{sgn} \left\{ \sum_{i=1}^n \alpha_i y_i K(x_i \cdot x) + b \right\}. \quad (7)$$

For these linearly non-separable data, a slack variable is added to obtain an optimum linear classifier. During our experiment, the RBF kernel $K(x_i \cdot x_j) = \exp \left\{ -\frac{|x_i - x_j|^2}{\sigma^2} \right\}$ are

selected with excellent performance for our pedestrian detection problem.

The proposed pedestrian detection system can be used on pedestrian detection problems on large images by the strategy of multi-scale sliding windows [2, 3]. The multi-scale sliding windows are used to extract pedestrian candidates with fixed step size. These sliding windows passed the validation of the proposed system are regarded as pedestrians of corresponding images. During these applications of pedestrian detections such as video monitoring and safe driving, these sliding window strategy can be added to the proposed pedestrian detection system.

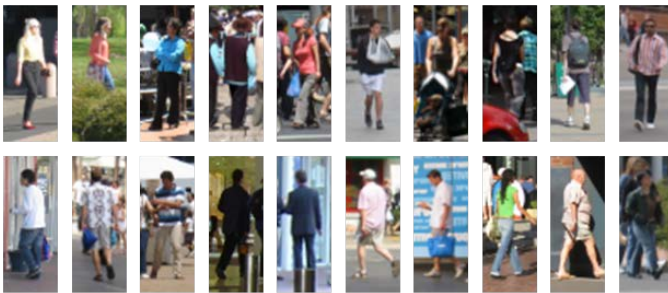
During pedestrian detection in large images, these single sample images extracted by multi-scale sliding are regarded as pedestrian candidates. We mainly focus on the verification of these single sample images to be pedestrians or not during our experiment simulation. This verification of single sample images are one of these necessary procedures of pedestrian detection. Therefore, the verification precision relates directly to pedestrian detection results in real applications.

3 Experimental Results and Discussion

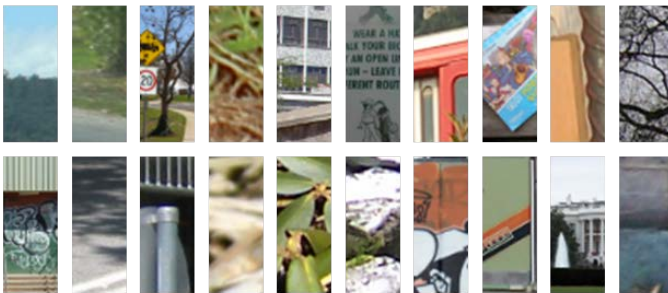
We evaluate the proposed pedestrian detection system based on SVM algorithm and LSH feature using five pedestrian detection datasets with different image scales from NICTA database [19]. The sizes of pedestrian and non-pedestrian images for these five datasets are $8 \times 20, 16 \times 20, 16 \times 40, 32 \times 40, 32 \times 80$. Partial pedestrian and non-pedestrian samples are demonstrated in Figure 1. The numbers of pedestrian and non-pedestrian samples are same and set to be 1000. All the experiments are carried out in

MATLAB R2013a environment running on a desktop with CPU 3.30 GHz and 16 GB RAM.

Aiming to validate the training and testing ability of the proposed system sufficiently, we use K-Fold Cross Validation technique [20] and $K=10$ in our experiment. During our experiment, LSH feature is compared with RHL, LBP and HOG feature, which is based on SVM classifier. For SVM classifier, we use L2 soft-margin method of RBF kernel with $\sigma = 2$. These parameters are selected through comprehensive comparison. Before trained by SVM classifier, these features are extracted and normalized to be $[-1, 1]$. For the LSH feature, the total number of bins is set to be $B=32$. The illumination invariant features of LSH feature are divided into 4×4 blocks with same scale size, and then the normalized summation of each block cascaded to be the input vector. We random select 85 dimensions of Haar-like features as the input vector of RHL feature. For the HOG feature, the gradients are voting into 9 orientation bins in $(0, \pi)$. During the calculation of LBP feature, the radius of each pixel is set to be 3 with surround 24 neighbours, and this feature is uniformly divided into 26 bins. The training / testing accuracies and times of these five datasets are demonstrated in Table 1.



Pedestrian samples



Non-pedestrian samples

Figure 1. Partial sample images from NICTA Pedestrian Dataset

The results in Table 1 show that training and testing times of LSH feature are similar as RHL, HOG and LBP features on these five dataset, based on SVM classifier. However, the overall performances training and testing accuracies of LSH feature are rather better than RHL, HOG and LBP features on these five datasets.

For all these five datasets, the LSH feature works better than RHL and HOG features. The training accuracies of LBP features are rather precise and always higher than 96% for all these five datasets. However, the testing accuracies are much

poor, and even less than 80% for 8×20 and 16×20 dataset. For other datasets with large image scales, the LBP has better performance than these datasets with small image scales. The RHL feature works well on these datasets with small image scales, such as 8×20 and 16×20 dataset, and much poor on other datasets with large image scales. The LSH and HOG features have more robust performances with image scale variation, and LSH feature is even better.

Table 1. Feature performance comparison with SVM classifier

Dataset	Featur e	Training		Testing	
		Accuracy(%)	Tim e	Accuracy(%)	Tim e
8×20	RHL	98.2	0.61	84.8	0.028
	HOG	87.2	0.56	82.5	0.013
	LBP	96.9	0.57	76.9	0.011
	LSH	94.9	0.58	89.1	0.019
16×20	RHL	96.5	0.62	80.7	0.036
	HOG	88.5	0.60	84.6	0.027
	LBP	98.0	0.56	78.8	0.020
	LSH	93.4	0.62	87.5	0.019
16×40	RHL	97.5	0.60	84.4	0.022
	HOG	87.6	0.56	84.0	0.028
	LBP	98.1	0.56	82.3	0.025
	LSH	95.0	0.58	90.0	0.023
32×40	RHL	92.1	0.60	76.1	0.033
	HOG	89.8	0.54	88.0	0.017
	LBP	97.4	0.57	86.2	0.027
	LSH	94.2	0.58	89.0	0.016
32×80	RHL	87.1	0.56	72.9	0.017
	HOG	89.7	0.56	87.7	0.022
	LBP	97.5	0.58	89.7	0.014
	LSH	95.1	0.57	89.9	0.016

4 Conclusions

In this paper, we proposed a novel pedestrian detection system with LSH feature based on SVM classifier. This LSH feature is robust with illumination variation and efficient to extract with integral image strategy. Compared with RHL, HOG and LBP features, LSH feature has excellent performance on these five datasets with better training and testing accuracies. SVM classifier is power and robust on these pedestrian datasets with parameters, which has been proved in our experiment. This proposed pedestrian detection system can be future utilized in the field of pedestrian detection on video monitoring and safe driving with multi-scale sliding window strategy on each frame. The pedestrian detection accuracies can be further enhanced with modified image features and machine learning algorithms.

Acknowledgment

This work is supported by the Basic Science Research Program through the Brain Korea 21 PLUS project and the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2013R1A1A2013778).

5 References

- [1] A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: Single-frame classification and system level performance." pp. 1-6.
- [2] L. Guo, P.-S. Ge, M.-H. Zhang, L.-H. Li, and Y.-B. Zhao, "Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4274-4286, 2012.
- [3] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 743-761, 2012.
- [4] H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hebert, and X. P. Maldague, "Advanced surveillance systems: combining video and thermal imagery for pedestrian detection." pp. 506-515.
- [5] O. L. Junior, D. Delgado, V. Gonçalves, and U. Nunes, "Trainable classifier-fusion schemes: an application to pedestrian detection."
- [6] N. Cristianini, and J. Shawe-Taylor, "An introduction to support vector machines," Cambridge University Press Cambridge, United Kingdom, 2000.
- [7] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, and T. Van Gestel, *Least squares support vector machines*: World Scientific, 2002.
- [8] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," *Pattern Recognition*, pp. 297-304: Springer, 2003.
- [9] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features." pp. I-511-I-518 vol. 1.
- [10] D. Tang, Y. Liu, and T.-K. Kim, "Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates." pp. 1-11.
- [11] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [12] P. Dollár, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification." pp. 1-8.
- [13] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," *Computer vision-eccv 2004*, pp. 469-481: Springer, 2004.
- [14] S. He, Q. Yang, R. W. Lau, J. Wang, and M.-H. Yang, "Visual Tracking via Locality Sensitive Histograms." pp. 2427-2434.
- [15] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," *Computer Vision-ECCV 2012*, pp. 864-877: Springer, 2012.
- [16] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, no. 18, pp. 3823-3831, 2011.
- [17] D. Wang, H. Lu, and M.-H. Yang, "Least Soft-threshold Squares Tracking," in Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, USA, 2013.
- [18] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 5, pp. 564-577, 2003.
- [19] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning." pp. 373-378.
- [20] D. Anguita, A. Ghio, S. Ridella, and D. Sterpi, "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines." pp. 291-297.

SESSION
STEREO, 3D IMAGING, DEPTH ALGORITHMS,
AND APPLICATIONS

Chair(s)

TBA

Depth Estimation of A Monocular Image from Image Blur

Zhonghai Deng¹, Jingyuan Zhang¹

¹Computer Science Dept. University of Alabama Tuscaloosa

Abstract— *In computer vision, estimating the depth information of object from an image/video has always been a hot topic, and finds many applications. In this paper, we make use of the image blurs caused by defocus aberration to estimate the relative distance in an object. To be more computationally efficient, we focus on the blurs along image edges. Experiments shows the effectiveness of our proposed method.*

1. Introduction

In computer vision, one essential problem is to reconstruct the 3-D scene from images. This could be simplified to estimate the distance of an object from captured still image. However, due to the information loss from 3D scene into 2D color image, this task seems un-solvable.

Robert *etc.* [1] uses the blur to analysis its affect on the perceived distance and size. They propose a probability model to help us understand how the blur indicates the apparent scale of the image's contents. Based on this assumption, they use a Bayesian' model to understand the distance. However, their result are judged, and compared with the human viewers responds, thus cannot give a confident and convincing result.

Another way, to estimate the image depth is done by Saxena *etc.* [3], where they use the image structure to find the relative distance for each pixels. By using a Markov Random Field (MRF) model, the give a more accurate and visually more pleasing result.

In optics and camera industry, the traditional way to compute depth is taking two images, one is a sharp image of the scene taken with a larger depth of field, while the other is blurred with varying the focal length.

One solution require using one image as a benchmark, and estimate blurs from the other one in various image points and compute its depth map/image of the scene accordingly. However in most of the times, people hardly have the benchmark image. Even in the real life, instead of like a which is actually a monocular image due to the fact that digital camera has only one 'eye'.

The 'focal length' is the distance between the lens plane and the sensor plane which is usually stated in millimeters . For cameras with zoom lenses, both the minimum and maximum focal lengths would be provided in the EXIF header file, for example 18Å\$55 mm. Though sometimes, they would be replaced by the Depth of Field (DOF) , indicating the distance range that objects could be sharp

enough. Another property of the camera is the angle of view, which is the visible area of the scene captured by the lens, stated as an angle. Wide angle of views captures larger areas, small angles gets smaller areas. Also, changing the focal length would automatically changes the angle of view.

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u} \quad (1)$$

Previously, researchers have used the defocus to compute depth or range, but mainly for the purpose of lens focus adjustment. [1] used the gradient in blur to compute range. [2] generalized it and extended the application of the defocus information for range computing while p241-cho present a matrix based method to estimate the camera shake blur. To compute depth from image blur, all these approaches utilize special features in scenes (or require texture in a small band of spatial frequencies). Also, all of them aim to adjust the focus of the camera, *i.e.* they use more than one image to do the range estimation, and adjust the camera settings accordingly.

$$e = (s, f, D) \quad (2)$$

if we define

$$q = \frac{2R}{D} = \frac{s-v}{v} = s \left[\frac{1}{v} - \frac{1}{s} \right] \quad (3)$$

Our method works in the opposite way. By revisiting the current problem, we propose a depth re-estimation based on an assumption that all image blur comes from the lens ill-focus. The method does not place any restrictions on the scene other than what is normal for passive depth imaging methods, *i.e.* we do not require that the scene must contain some texture or other visual features. For any method lack of features or texture should cause graceful degradation, however. We introduce a confidence which indicates the reliability of the computed depth estimates. Actually, we believe that the use of a confidence measure is a key point for the usefulness of the idea.

2. Blur Measurement

In today's commercial digital camera, the most important part is the sensor, which transfer the light signals into the electric ones. Before that, camera need a lens system to focus the lights onto the sensor. To better control this focusing process, the lens system has four degrees of freedom: shutter speed (exposure time to sensor), focus (position the sensor should be), focal length (zoom), and aperture (diameter of the lens cover).

For simplicity, we assume an optical system is only a thin convex lens (see Figure 1). Here, the lights coming from the out-world scene, are focused on to the light detector (sensor). We use the f for the focal length, s for the distance from the sensor to the lens, v for the distance from the lens to the focal plane. The scene object is placed at a distance u from the lens.

From Figure 1, we can see that sometimes, when the lights are focused on position p' , the image plane is positioned behind, thus the lights coming from p are longer focused on a single point on the image plane, forming a blurred circle, with diameter R . The similar case when the object plane is closer to the lens than the focal plane p' .

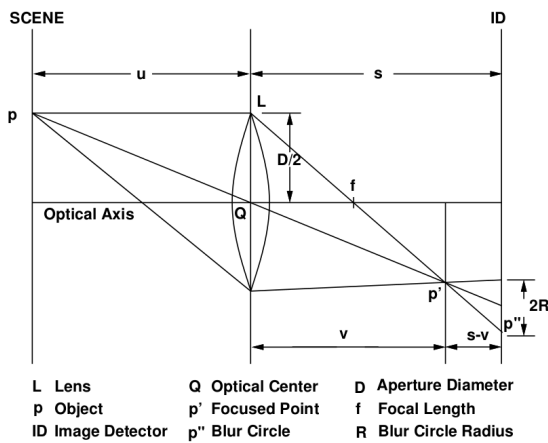


Fig. 1: *Imaging Formation in a Convex Lens (Reproduced from [4])*

We know that the lens can only focus at a certain point, therefore, when recording images of 3D scenes we have the situation in Figure 1, the objects in the scene appear more or less blurred on sensor. Thus many researchers take two consecutive images, and measure the change in blur from the first image to the next we may compute the depth.

However, in our setting, we have only one image available, *i.e.* without the benchmark image, we can only use this monocular image for depth estimation.

First, we need find the image part at focus, and by extracting the image header EXIF file of that part, we can estimate the image depth, and use that blur as the background image noise. In this way, we can estimate its relative range with regard to the distance at focus.

In our experiments, instead of estimating every pixel point for each pixel, in this paper, we try to estimate the absolute image distance for each image object. And this distance within objects are interpolated for computational consideration.

2.1 Blur Model

An edge is modeled as a step function $Au(x) + B$ of unknown amplitude A and pedestal offset B , which, for the purposes of this discussion, will be aligned with the y -axis of the image coordinate frame. The focal or penumbra blur of this edge is modeled by the Gaussian blurring kernel

$$g(x, y, \lambda_b) = \frac{1}{2\pi\lambda_b^2} e^{-frac{(x^2+y^2)}{2\lambda_b^2}} \quad (4)$$

where $g(x, y)$ is the pixel value at an image, λ_b is the distance constant factor that generate the blur function.

Estimating the sensor noise statistics for an imaging system is relatively straightforward. For the system used in Fig. 2, a region of a de-focused image of a plain flat surface was first selected.

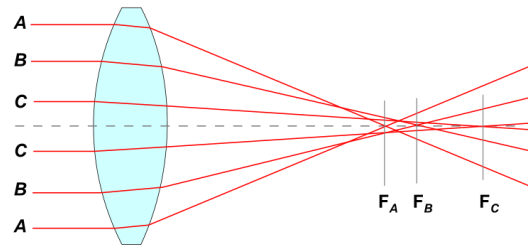


Fig. 2: *Imaging Blur Type due to Imperfection Focusing*

Our model predicts that the camera imaging system estimates absolute distance of the object at focus by finding the depth of focus from header file.

three types of blur: (1) blur that is completely consistent with the relative distances in a scene (consistent-blur condition),

Thus, to simplify our computation, and get a more smooth result, we assume that in the perfect condition, all edges should be sharp, *i.e.* edges of objects at focus should be only one pixel size. (In real world, perfect edge should be of no width theoretically.)

3. Depth Estimation via Image Blur

Theoretically speaking, for each pixel values of an image, there exist an depth value. However, due to the various types of constrains, it is impossible, and unnecessary to do the depth estimation for each pixel.

So, in this paper, we first segment the image into different objects, and estimate these value according to their focused point. And then, we estimate objects in other parts to access their depth value. We do it by using the blurs of their contouring edges.

3.1 Extract Image Focal Length

Every image file from a digital camera comes naturally with a header file. Though its format varies from brand to

brand, model to model, it generally contains some basic information, like image file name, image size, number of colors, file size, file type and so on. Another feature that must be included is the 'Focal Length', which usually use 'mm' as the measurement unit.

In our experiments, we use the ExifTool to extract the this focal length. There are two other measures that we are also interested. One is the F number, meaning the $\frac{f}{D}$ value. The other value is the image depth of field (DOF) value of an image, though not

In optics, DOF is the distance between the nearest and farthest objects in a scene that appear acceptably sharp in an image. Although a lens can precisely focus at only one distance at a time, the decrease in sharpness is gradual on each side of the focused distance, so that within the DOF, the unsharpness is imperceptible under normal viewing conditions.

3.2 Inference Distance from Blur

So, we first expose the image part, or objects that is on focus, which is achieved by comparing the sharpness metric inside each image part.

And then, the absolute depth value could be calculated as

$$u = \frac{fv}{v-f} \quad (5)$$

Where f is the focal length, which can be extracted from the image EXIF header. And from figure 1 we have

$$\frac{2R}{D} = \frac{S}{v} \quad (6)$$

Where $D = \frac{f}{F_{num}}$, and F_{num} could also be extracted from the EXIF info. Thus we have

$$v = s * \frac{2R + D}{D} \quad (7)$$

As we are only interested in the relative depth of the image, *i.e.* the absolute value of the object at focus is not needed. Therefore, in the following sections, we shrink the value s , so that distance of the object at focus is set to be 1, and other distances are compared with it.

From equation 3, we can see the relationship between camera optics and image blur. Also, the human visual system uses of retinal-image blur to adjust the focal length. For instance, if our target object is blurred, our pupils would automatically adjust the focal length so that the focused point would fall onto the retina.

From above figure, we can see that the more blur occurs, the larger radius R projected on the sensor plane.

Also, these blurs contain other depth cues if we can determine the objects at focus, and compare them with the rest. (*i.e.* linear perspective, relative size, *etc.*). This information help to specify the relative distances among objects in the scene. Although this distance are scale ambiguous, thus we seems cannot directly calculate the absolute distances of the

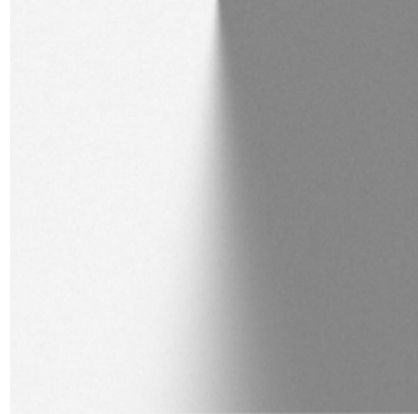


Fig. 3: Imaging Edge width due to Blurs

interested objects. We can actually, use the f number to help determine this value.

4. Experimental Result and Analysis

In this part, we apply our algorithms on the images taken by commercial cameras. However, to be more comparable with current depth estimation result, we use the well-known peppers image.

4.1 Determine the Beacon Distance of Object at Focus

From the image EXIF header, we can always extract the image depth of field (DOF) value of an image, as well as the focal length. Also, there is also a F number, which means the $\frac{f}{D}$ value.

So, we first expose the image part, or objects that is on focus, which is achieved by comparing the sharpness metric inside each image part.

And then, the absolute depth value could be calculated as

$$u = \frac{fv}{v-f} \quad (8)$$

Where f is the focal length, which can be extracted from the image EXIF header. And from figure 1 we have

$$\frac{2R}{D} = \frac{S}{v} \quad (9)$$

Where $D = \frac{f}{F_{num}}$, and F_{num} could also be extracted from the EXIF info. Thus we have

$$v = s * \frac{2R + D}{D} \quad (10)$$

As we are only interested in the relative depth of the image, *i.e.* the absolute value of the object at focus is not needed. Therefore, in the following sections, we shrink the value s , so that distance of the object at focus is set to be 1, and other distances are compared with it.

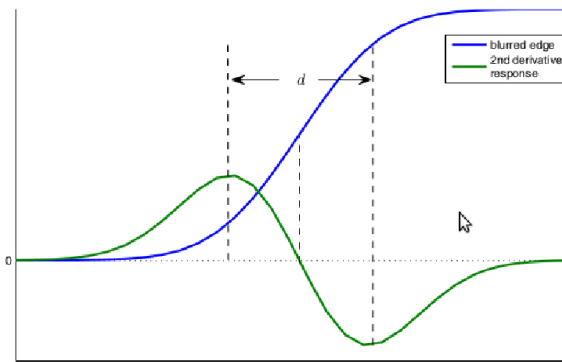


Fig. 4: Image Blur Edge Derivative

4.2 Determine the Beacon Distance of Object out of Focus

In our experiments, we set only one object at focus, and the rest would be depth estimated from their blur value. For computational consideration, we infer the depth for the edges, and the inner point would be interpolated from edges.

Due to the fact that traditional edge finding algorithms tends not to find a closed area for an object, we use the contour detection and hierarchical segmentation algorithm borrowed from ??, to find the boundary, and estimates how much blurs along the boundary.

4.3 Analysis of Detection

Due to the fact that we do not have the ground truth depth map, we only print the depth information from the image. But this depth estimation method produce a convincing result on its object depth estimation. However, our method does not perform well on the case that for image parts out of DOF (Depth of Field), it would produce a unreliable result. Also, no bench-mark database could be used to gain the comparison with other methods.

5. Conclusions and Future Work

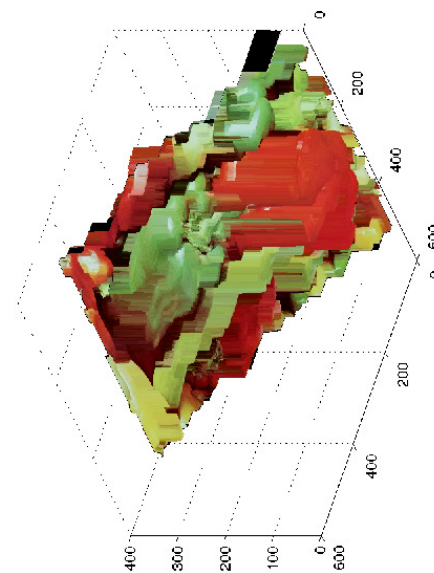
In this paper, we propose a novel image blur estimation method, and estimate the depth from a single blurred image. We have also reviewed the achievements, discuss the experimental result . Even though our proposed method can give an exciting estimation result, the depth maps it produces are not always satisfying. Also, one drawback of our proposed method is that it cannot differ the blurs caused by before/behind focal plane (*i.e.* too near, and too far from focused point). This problem might be solved by combining other methods, like structure inference.

References

- [1] R.T. Held, E.A. Cooper, J.F. O'BRIEN, and M.S. Banks. Using blur to affect perceived distance and size. *ACM transactions on graphics*, 29(2), 2010.



(a) Blurred Edge



(b) Amplitude model of the blurred edge, and its second derivative

Fig. 5: Blurred Edge, horizontal intensity and its 2^{nd} derivative.

- [2] P. Kakar, N. Sudha, and W. Ser. Exposing digital image forgeries by detecting discrepancies in motion blur. *Multimedia, IEEE Transactions on*, 13(3):443–452, june 2011.
- [3] A. Saxena, M. Sun, and A.Y. Ng. Make3d: Learning 3d scene structure from a single still image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):824–840, 2009.
- [4] Murali Subbarao and Gopal Surya. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, 13:271–294, 1994.

Optimum Image Quality Assessment for 3D Perception of Stereoscopic Image Generated from Upsampled Depth Map

Saeed Mahmoudpour and Manbae Kim

Dept. of Computer and Communications Engineering, Kangwon National University
Chunchon, Republic of Korea

E-mail: {saeed, manbae}@kangwon.ac.kr

Abstract –Depth map upsampling is an approach to increase the spatial resolution of depth maps obtained from ToF (Time of Flight) cameras. Since depth map quality directly affects 3D perception of stereoscopic image, applying different depth upsampling methods to a low resolution depth map causes a variety of perceptions of the stereoscopic images. In this paper, we investigate the relation between objective measurements and 3D subjective evaluation. For the former, diverse full-reference and no-reference quality assessment tools are applied to measure the quality of depth maps obtained. Subjective evaluation is carried out by DSCQS test on the stereoscopic images. We utilize several upsampling methods to achieve different depth map qualities of a scene as our experiment samples. Finally, the quantitative value of each measurement is compared with a subjective assessment using three correlation coefficients. The experiment shows promising results that could help to select the most appropriate objective quality assessment tool(s) for stereoscopic image quality measurement.

Keywords: Depth map, Upsampling, Objective assessment, Correlation

1 Introduction

Following the recent technology advances in camera systems and computer vision, three-dimensional active cameras are capable to provide accurate distance information of a scene. The high speed Time of Flight (ToF) cameras extract reliable depth maps. However, the spatial resolution of depth maps is relatively low in comparison with original images. Therefore, diverse depth map upsampling approaches are provided to obtain high-resolution depth maps. Also, it is important to evaluate the upsampling quality in order to realize upsampling performance on 3D content quality.

Image quality assessment (IQA) helps us to evaluate the quality of the upsampled depth map using any assessment tools. The most frequently used method is PSNR (Peak Signal-to-Noise). As well, due to various sources of quality degradation and visual discomfort which degrade the end-user 3D experience and lack of accurate objective IQA tools, the

subjective assessment is commonly used for 3D quality evaluation.

Investigating the similarity between 2D quality evaluation and 3D perception, we will search for the most accurate IQA tool for 3D quality assessment. Using the proper automatic objective IQA tool will help to predict the quality of 3D image without using the expensive subjective test and even without watching the stereoscopic image.

Depth map quality has significant effect on 3D perception, therefore, first we have implemented seven upsampling methods and the depth map quality of each method is computed using different objective IQA tools. Then, the quality results are compared with a reliable subjective assessment using correlation.

The Pearson, Spearman and Kendall correlations are three proper approaches for similarity measurement between each objective assessment result and the subjective IQA. Comparing the objective and subjective scores, it will be inferred that which objective IQA tools show the most correspondence with human judgment and have superiority for 3D quality assessment. DIBR (depth image based rendering) or 2D+Depth is used to generate a stereoscopic image. Fig. 1 shows the overall framework that examines the relation between upsampling methods and 3D perception.

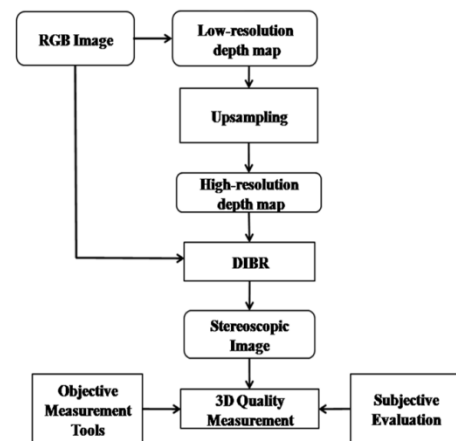


Fig. 1. The flow diagram for testing 3D quality of depth upsampling methods.

Since it is difficult to investigate all methods, seven approaches are selected to be used in this work. The *bilinear upsampling* (BLU) uses average weighted of four neighboring pixels for interpolation to achieve upsampled depth map. A similar method called *bicubic upsampling* (BCU) is based on sixteen neighboring pixels. The *bilateral upsampling* (BU) [1] is a prevalent approach combines a spatial filter and a range filter to preserve the edge regions in upsampling process. Another upsampling method based on the bilateral upsampling is *joint bilateral upsampling* (JBU) [2] which utilizes both a color data and its low-resolution depth map. The *variance-based upsampling* (VBU) [3] avoids the usage of the constant variance and uses a variance that is computed on each pixel block. The disadvantage of the JBU is that it is sensitive to homogeneous regions and the weighting function can be assigned a wrong variance in non-edge regions. To solve this problem, an *adaptive bilateral upsampling* method (ABU) [4] has been proposed, where a large weight is assigned to color image at edge pixels and a large weight is assigned to depth data at non-edge pixels. To overcome the limitation in reducing blur at low-gradient edge regions in prior methods, a *distance transform-based bilateral upsampling* (DTBU) [5] has been proposed.

This paper is organized as follows. In the next section, different IQA tools considered in this work are described. The correlation coefficients for similarity measurements are introduced in section 3 and the experimental results with intensive discussion are provided in section 4. Finally, we summarize our work in section 5.

2 Objective Quality Assessment Tools

Full-reference image quality assessment (FR IQA) compares test and reference images, therefore, both ground-truth and upsampled depth map are needed. The no-reference/blind image quality assessment (NR IQA) refers to quality assessment of images by an algorithm where only the distorted image is accessible and no information about the reference image is available. In this paper, several FR IQA and NR IQA tools are used to evaluate the performance of different upsampling methods. The quality metrics are introduced as follows:

2.1 FR IQA tools

A. PSNR

PSNR is one of the most prevalent tools for image quality evaluation defined by (1):

$$PSNR = 10 \cdot \log \left[\frac{\sum (D^h - D^u)^2}{255^2} \right] \quad (1)$$

where D^h and D^u are ground-truth and upsampled depth maps respectively.

B. SSIM

A sophisticated tool for image quality evaluation is structural similarity index measure (SSIM) [6] that measures the similarity between two images and considered to be correlated with the quality perception of the human visual system (HVS). SSIM principle is based on the modeling any image distortion as a combination of luminance distortion, contrast distortion and loss of correlation. SSIM value for two images f and g is expressed by

$$SSIM = l(f, g)c(f, g)s(f, g) \quad (2)$$

$$l(f, g) = \frac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1}, c(f, g) = \frac{2\sigma_f\sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2}, s(f, g) = \frac{\sigma_{fg} + C_3}{\sigma_f + \sigma_g + C_3}$$

where $l(f, g)$, $c(f, g)$ and $s(f, g)$ are luminance, contrast and structure comparison functions respectively. σ_f and σ_g denote standard deviations, μ_f and μ_g are mean values and σ_{fg} is covariance. C_1 , C_2 and C_3 are positive constants added to avoid a null denominator. The SSIM is a value between 0 and 1 that higher value shows more similarity.

C. VIF

Visual Information Fidelity (VIF) [7] is a full-reference image quality metric that uses information theoretic criterion for image fidelity measurement. In an information-theoretic framework, the information that could ideally be extracted by the brain from the reference image and the loss of this information to the distortion are quantified in VIF method using natural scene statistics (NSS), HVS, and an image distortion (channel) model. The VIF is derived from a quantification of two mutual information quantities: the mutual information between the input and the output of the HVS channel when no distortion channel is present (called the *reference image information*) and the mutual information between the input of the distortion channel and the output of the HVS channel for the test image. Similar to SSIM, the assessment result is represented using a value between 0 and 1.

2.2 NR IQA tools

A. Sharpness Degree

Sharpness degree⁽⁸⁾ is used to represent the extent of sharpness of the image and is defined by (2).

$$\text{Sharpness Degree} = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} G^2(x, y) \quad (3)$$

where

$$G(x, y) = \sqrt{(D(x, y) - D(x-1, y))^2 + (D(x, y) - D(x, y-1))^2}$$

B. Blur Metric

Another tool for measuring blur attempts to measure the spread of the edges. First, we apply an edge detector (e.g. a Sobel edge detector) to a grayscale image. We scan each row of the image for pixels corresponding to an edge location. The start and end positions of the edge are defined as the locations of the local extrema closest to the edge. The spread of the edge is then given by the distance between the end and start positions, and is identified as the local blur measure for this

edge location. The global blur measure for the whole image is obtained by averaging the local depth values over all edges found [9].

$$\text{Blur Metric} = \frac{\text{Sum of all edge widths}}{\text{No. of edges}} \quad (4)$$

C. BIQI

Blind image quality index (BIQI) [10] identifies the likeliest distortion in the image and then quantifies this distortion using a NSS-based approach. Given a distorted image, the algorithm first estimates the presence of a set of distortions in the image that consists of JPEG, JPEG2000, white noise (WN), Gaussian Blur (Blur) and Fast fading (FF). The amount or probability of each distortion in the image is denoted as p_i $\{i=1,2,\dots,5\}$. The method performs quality assessment in two stages. This first stage is a classification and the second stage attempts to evaluate the quality of the image along each of these distortions. The quality of the image is then expressed as a probability-weighted summation:

$$\text{BIQI} = \sum_{i=1}^5 p_i \cdot q_i \quad (5)$$

where q_i $\{i=1,2,\dots,5\}$ represents the quality scores from each of the five quality assessment algorithms (corresponding to the five distortions).

D. NIQE

Natural image quality evaluator (NIQE) [11] is a completely blind image quality analyzer that only uses measurable deviations from statistical regularities observed in natural images, without training on human-rated distorted images. Unlike current general purpose no reference (NR) IQA algorithms which require knowledge about anticipated distortions in the form of training examples and corresponding human opinion scores, NIQE uses a quality aware collection of statistical features based on the simple and successful space domain, the NSS model. These features are derived from a corpus of natural, undistorted images.

The quality scores for both BIQI and NIQE are expressed by a value between 0 and 100 (0 represents the best and 100 the worst quality).

3 3D Subjective Quality Assessment

The quality of stereoscopic images made by DIBR technique is evaluated using subjective quality experiment. We observed the stereoscopic images with a 3D monitor adopting DSCQS (Double Stimulus Continuous Quality Scale) subjective test [12]. In the first stage, original views were displayed to ten participants. Each participant watched an original stereoscopic image for 10 seconds and another stereoscopic image made by an upsampled depth map for the same period, and the effect of the 3D depth is evaluated. For each image data, similar viewing was carried out in order to examine the 3D perception. Depth perception was then subjectively judged on a scale of 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent) in terms of 3D perception.

4 Correlation Measurement Metrics

In order to investigate the relationship between objective measurement tools and 3D subjective evaluation, three approaches are utilized.

4.1 Pearson Correlation A popular metric for measuring the association of two continuous variables is Pearson's correlation coefficient. Pearson method shows the strength of relationship using a coefficient ranges from -1 to 1. Positive coefficient means simultaneous changes in two variables, negative coefficient implies inverse association and zero means no association between variables. Pearson's coefficient (ρ) is defined by

$$\rho = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \quad (6)$$

where $\text{Cov}(x,y)$ is the covariance between two groups and σ_x and σ_y denotes standard deviations.

4.2 Spearman Correlation This correlation coefficient is a rank-based version of the Pearson's correlation coefficient. First, the samples of each group with n variables are ranked from 1 to n (value 1 shows the smallest and value n denotes the biggest sample). Then, Spearman's correlation coefficient (ρ_s) can be calculated as

$$\rho_s = \frac{\sum_{i=1}^n ((\text{rank}(x_i) - \overline{\text{rank}(x)}) (\text{rank}(y_i) - \overline{\text{rank}(y)}))}{\sqrt{\sum_{i=1}^n ((\text{rank}(x_i) - \overline{\text{rank}(x)})^2) \sum_{i=1}^n ((\text{rank}(y_i) - \overline{\text{rank}(y)})^2)}} \quad (7)$$

where $\text{rank}(x_i)$ and $\text{rank}(y_i)$ are the ranks of the observation in the sample. Similar to Pearson, Spearman's coefficient varies from -1 to +1 and the absolute value of ρ_s indicating the strength of association.

4.3 Kendall's Tau Correlation Kendall's tau ranges from -1 to +1 and describes the strength of the relationship between the two variable similar to previous correlation coefficients. This metric is defined to measure how much two variables are correlated. This coefficient quantifies the difference between the number of concordant and discordant pairs. Any two pairs of ranks (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant when $(x_i - x_j)(y_i - y_j) < 0$. Kendall's tau coefficient is

$$\tau = \frac{\sum_{i=1}^n \sum_{j=1}^n \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{n(n-1)} \quad (8)$$

where

$$\text{sgn}(X) = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{if } X = 0 \\ -1 & \text{if } X < 0 \end{cases}$$

5 Experimental Results

The quality performance of the seven upsampling methods are evaluated using ten test depth maps from Middlebury

stereo dataset [13]. The test RGB images and related depth maps are shown in Fig. 2. In order to obtain low-resolution depth maps, we downsampled the original data and then, we made high-resolution depth maps.

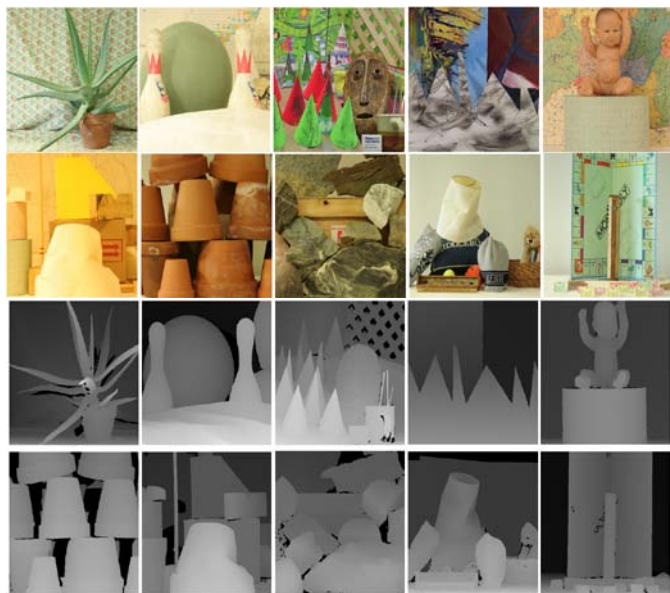


Fig. 2 Test RGB and depth maps provided by Middlebury [12]

Figs. 3~5 show the upsampled depth maps of the seven methods for *aloe*, *cone*, and *bowling*. The stereoscopic images are shown in Fig. 6.

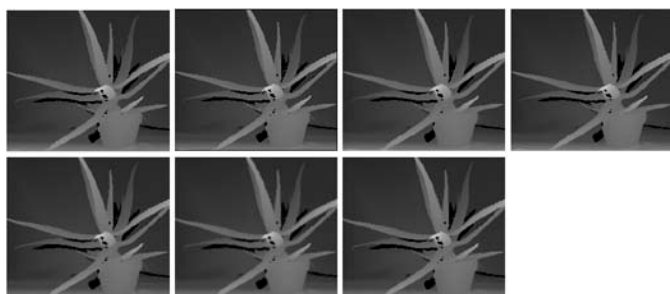


Fig. 3 Upsampled depth map of *aloe* obtained by BLU, BCU, BU, DTBU, JBF, ABU and VBU in the scan order.

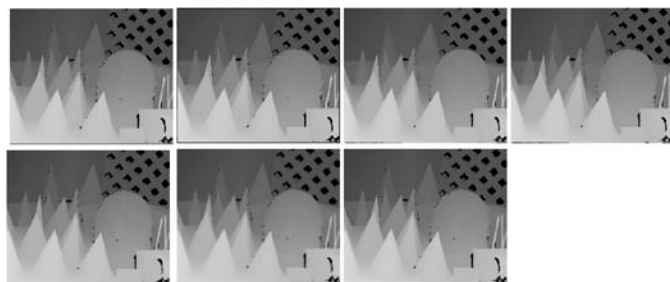


Fig. 4 Upsampled depth map of *cone* obtained by BLU, BCU, BU, DTBU, JBF, ABU and VBU in the scan order.

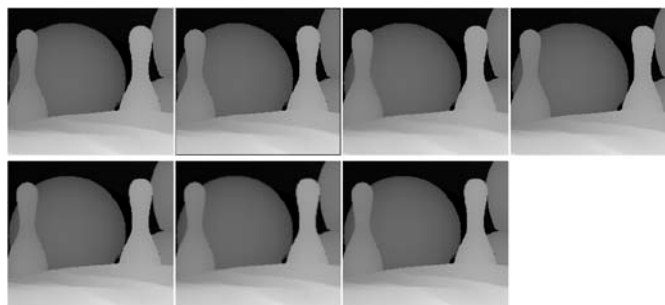


Fig. 5 Upsampled depth map of *bowling* obtained by BLU, BCU, BU, DTBU, JBF, ABU and VBU in the scan order.

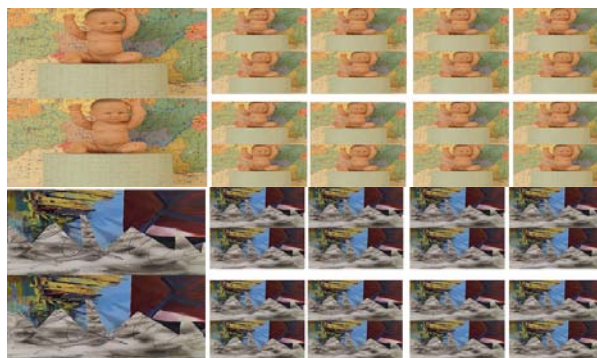


Fig. 6 Stereoscopic images in top-bottom format.

Table 1. Average subjective measurement data of upsampled depth maps

	BLU	BCU	BU	JBU	VBU	ABU	DTBU
Visual Fatigue	3.76	3.64	3.8	3.84	4.03	3.46	3.99

Table 2. Average objective measurement data of upsampled depth maps (PSNR unit: dB)

Depth map	BLU	BCU	BU	JBU	VBU	ABU	DTBU
PSNR image	35.85	35.71	35.64	34.15	35.64	33.16	34.86
Edge PSNR	23.68	23.55	23.66	22.82	23.38	20.97	22.93
Non-edge PSNR	38.07	37.94	37.78	37.50	37.93	35.43	36.92
Sharpness	39.5	42.2	49.51	49.09	31.92	88.31	68.14
Blur	8.48	11.38	10.29	10.87	10.51	9.00	9.89
SSIM	0.976	0.955	0.975	0.956	0.971	0.962	0.972
Edge SSIM	0.957	0.915	0.955	0.944	0.952	0.942	0.9
VIF	0.518	0.539	0.424	0.422	0.478	0.398	0.438
BIQI	57.8	66.34	63.11	32.81	41.94	29.15	72
NIQE	15.95	13.11	13.94	11.82	12.47	13.41	13.82

Table 1 and 2 represent the quality scores for each upsampling method measured from different objective quality metrics and subjective experiment (Visual Fatigue) respectively. The results are derived from averaging the quality scores of the collection of 16 images.

The 3D perception grades of each upsampling method in Table 1 are based on 3D visual fatigue. The visual fatigue values obtained by the subjective test are 3.76 (BLU), 3.64

(BCU), 3.89 (BU), 3.84 (JBF), 4.03 (VBU), 3.46 (ABU) and 3.99 (DTBU). Table 2 compares the objective measurement data of the seven upsampling methods.

The quality scores of upsampling depth maps obtained from each IQA metric is considered as a group of seven samples. All values are normalized by scaling between 0 and 1 and the similarity of samples distribution in each IQA group is compared with visual fatigue samples group using Pearson, Spearman and Kendall correlation coefficients. Table 3 shows the correlation results.

Before evaluating the strength of correlation using different correlation coefficients, it is worth mentioning that Pearson's correlation coefficient takes into account both the number and degree of concordances and discordances, whereas Kendall's tau correlation coefficient shows only the number of concordances and discordances. Spearman's correlation is in between of the Pearson's and Kendall's, reflecting the degree of concordances and discordances on the rank scale. The disadvantage of Pearson is the sensitivity to outliers (an observation that is numerically distant from the rest of the data). In this case, Spearman and Kendall are less sensitive to outliers and preferable.

According to Table 3, edge PSNR shows higher value of correlation compare to common PSNR and non-edge PSNR. Also, Pearson coefficient is much higher than Spearman result that indicates the distribution is non-linear. In this case, Spearman and Kendall results are more reliable.

Sharpness degree and blur metric show negative and positive correlation values respectively. These two results confirm the fact that image with high spatial frequency (sharper) reveals much noticeable visual discomfort than that with low frequency [14].

SSIM uses luminance, contrast and structure features to measure quality. Similar to PSNR, Pearson coefficient is higher than two other correlation coefficients in this metric. SSIM has the highest Spearman value among other metrics. Thus, it is the most similar metric to visual fatigue in the case of samples order.

As mentioned earlier, Pearson is very sensitive to outliers and its value can be drastically influenced by a few extreme values. Negative value of Pearson with positive Spearman and Kendall coefficients for edge-SSIM show that Pearson can not be used due to outliers, therefore, Pearson correlation may severely understate the strength of a relationship between two variables. In this case, we should rely on Spearman results that reveals correlation more than edge-PSNR but less than SSIM.

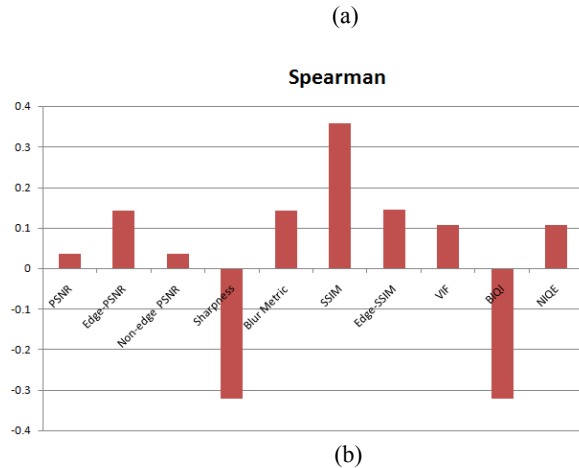
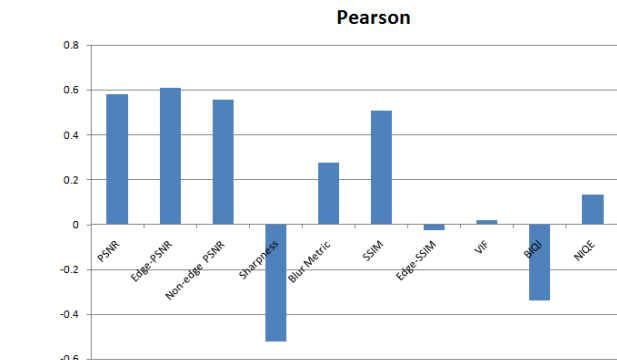
VIF results are based on NSS, HVS, and an image distortion (channel) model in wavelet domain and shows a positive but low correlation to visual fatigue.

BIQI and NIQE are two NR IQA metrics that are expected to show lower correlation in comparison to FR IQA metrics. JPEG, JPEG2000 (JP2K), white noise (WN), Gaussian Blur (Blur) and Fast fading (FF) are five distortions that are considered in BIQI method for quality measurements. Similar to negative results of Sharpness degree, BIQI is not correlated with visual fatigue results.

NIQE metric delivers a positive correlation using Pearson coefficient. Also, Spearman and Kendall correlation results are comparative to some results derived from FR IQA methods. NIQE results are near to VIF, therefore, it can be inferred that NIQE is an acceptable quality assessment tool when there is no access to reference image. Fig. 4 shows correlation values for different quality metrics in column diagram mode.

Table 3. Pearson, Spearman and Kendall correlation coefficients between visual fatigue and objective measurements

	Pearson	Spearman	Kendall
PSNR	0.582	0.0357	0.0476
Edge-PSNR	0.608	0.1429	0.1429
Non-edge PSNR	0.554	0.0357	0.0476
Sharpness	-0.522	-0.3214	-0.1429
Blur Metric	0.273	0.1429	0.0476
SSIM	0.505	0.3571	0.1429
Edge-SSIM	-0.025	0.1441	0.0476
VIF	0.019	0.1071	0.1429
BIQI	-0.339	-0.3214	-0.2381
NIQE	0.132	0.1071	0.1429



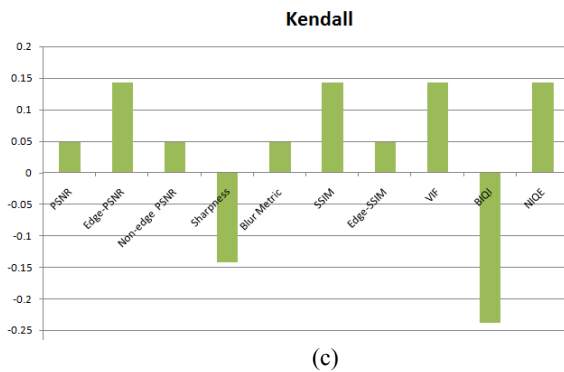


Fig. 7. Column diagram of correlation between image quality metrics and visual fatigue in descending order using (a) Pearson, (b) Spearman and (c) Kendall

6 Conclusions

In this paper, we investigated a method to find the superior quality assessment tools, which are commonly used in 2D images, for 3D quality assessment. Using correlation, we successfully achieved a reasonable relation between objective QA results and subjective assessment. Also, several upsampling results are provided to use as group samples in this work. As a result, PSNR and SSIM show the highest Pearson correlation coefficients. However, corresponding Spearman and Kendall for PSNR assessments are far from Pearson coefficients. The reason of this difference is outliers or highly skewed variables, therefore, the distribution is not linear and Pearson is unreliable.

According to Spearman and Kendall coefficients for Edge-PSNR, this metric is less correlated compare to SSIM but better than other metrics. Consequently, SSIM is the most concordant metric according to Spearman coefficient and also with high Pearson and Kendall coefficient values. Furthermore, Sharpness Degree and BIQI are not appropriate tools due to their negative correlation coefficients.

7 Acknowledgements

This work was supported by Small & Medium Business Administration (SMBA) under Technology Innovation Project (S2056930) and in part by the MKE (The Ministry of Knowledge Economy)/KEIT [10041082, System and Semiconductor Application Promotion Project].

8 References

- [1] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Image," In Proc. IEEE Int. Conf. on Computer Vision, pp. 836-846 (1998).
- [2] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, "Joint bilateral upsampling," ACM Trans. on Graphics, Vol. 26, No. 3, pp.1-6 (2007).

- [3] S. Jang, D. Lee, S. Kim, H. Choi, M. Kim, "Depth Map Upsampling with Improved Sharpness," Journal of Broadcast Engineering, Vol. 17, No. 6, pp. 933-944 (2012).
- [4] C. Pham, S. Ha, and J. Jeon, "A local variance-based bilateral filtering for artifact-free detail- and edge-preserving smoothing," PSIVT, Part II, LNCS 7088, pp. 60-70 (2011).
- [5] D. Yeo, E. Haq, J. Kim, M. Baig, H. Shin, "Adaptive Bilateral Filtering for Noise Removal in Depth Upsampling," SoC Design Conf., pp. 36-39 (2010).
- [6] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", IEEE Trans. On Image Processing, Vol. 13, No. 4 (2004).
- [7] H.R. Sheikh and A. C. Bovik, "Image information and image quality", IEEE Trans. On Image Processing, Vol. 12, No. 2, pp. 430-444 (2006).
- [8] C. Tsai, H. Liu, M. Tasi, "Design of a scan converter using the cubic convolution interpolation with canny edge detection" International Conference on Electric Information and Control Engineering (ICEICE), pp. 5813-5816, (2011).
- [9] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to JPEG2000,"Int. Workshop on Multimedia Signal Processing, pp.403-408 (2008).
- [10] A. K. Moorthy and A. Bovik, "A two-step framework for constructing blind image quality assessment", IEEE Signal Processing Letters, Vol. 17, No. 5, pp. 513-516, (2010).
- [11] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer", *IEEE Signal Processing Letters*. Vol. 20, No. 3, pp. 209-212 (2013).
- [12] E. Lee, H. Heo, and K. Park, "The comparative measurements of eyestrain caused by 2D and 3D displays," IEEE Trans. on Consumer Electronics, Vol. 56, No. 3, pp. 1677-1683 (2010).
- [13] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," Int. J. of Computer Vision, Vol. 47, No. 1-3, pp. 7-42, 2002.
- [14] D. Kim and K. Sohn, " Visual fatigue prediction for stereoscopic image", IEEE Trans. Circuits Syst. Video Technol., Vol. 21, No. 3, pp. 231-236 (2011).

A Hole-Filling Approach Based on Morphological Operation for Stereoscopic Image Generation

Chyuan-Huei Thomas Yang, Yen-Ting Hsu, Wen-Chieh, Chou

Department of Information Management, Hsuan Chuang University, Hsin Chu, Taiwan

Abstract - Apply the stereoscopic image generation from 2D to 3D conversion depth image based rendering (DIBR) technique it is 2D view to multi-angle virtual view by the single depth image. We develop a stereoscopic image generation method from one-view color image and its corresponding depth information. The DIBR firstly does the preprocessing of the depth image by using the smooth filter, which sharpens the discontinuous depth changes as well as to smooth the neighboring depth of similar color and also detain noises from appearing on the warped images. The occlusion regions are applied the morphological operations on the depth image with the background depth levels to keep the depth structure. With depth-guided exemplar-based image inpainting combines the color gradient to preserve the image structure in the restored regions.

Keywords: Depth image based rendering, Hole filling approach, Color gradient, Morphological operations

1 Introduction

3D Television has become an important entertainment medium, but in recent 3D contents of a large number of acceptable not to be one of the reasons is the lack of, and often is not currently suitable for initial purchase 3D TV causes. A quick fix the lack of 3D contents programs, namely 3D broadcast on television by the 2D transform 3D programs or movies. Today a number of 3D TVs already have this automatic conversion. However, the 2D-3D conversion, can only offer simple simulation 3D rather than genuine 3D. Actually it is not really 2D transformed into actual 3D, or just adding an actual depth textures within the image. In recent years, people will want to experience the more 2D images more realistic 3D Visual effects, 3D related technologies were becoming increasingly concerned by people, and stereoscopic display technology has greatly improved. Generation of stereo images can be divided into two broad categories: (1) by the dual-lens stereoscopic images; (2) single-camera images converted to multiple perspectives of the virtual image. Current twin-lens 3D camera is expensive and a twin-lens camera calibration and synchronization issues, internal parameter adjustment is also very complex. Mostly owned by the photographer is still dominated by single lens camera or camera. So many experts and scholars in relevant fields are studying or applying 2D images into 3D images, so we may avoid buying the expensive 3D Cameras and accompanied by undesired operation problem, also we can make 3D images rich and popular.

Internationally well-known seminar (Such as SIGGRAPH, and CVPR) stereoscopic images made by the relevant technical and learned that in computer graphics and computer vision technology has begun to integrate, wants to provide a more realistic and more in stereoscopic images. With the future of stereoscopic display technologies in recent years has become more mature, interactive full view 3D TV (interactive free-viewpoint 3D TV) household TV will be the next generation of new specifications, the showing of the film will not just be mere stereo image, will be further high quality 3D Stereo images of the model. This program is a three-dimensional image synthesis techniques will be explored, can be divided into two broad classes, image drawing method (Image based rendering,IBR), as well as the depth of image drawing method (Depth image based rendering,DIBR), which don't use depth information-image drawing method, but because of the restrictions on use, it is less frequently used. Current depth image drawing method for 2D conversion to 3D image synthesis technology in their evolution that adds up the stereo to stereo system for object image synthesis, mainly includes three parts: (1) pretreatment (pre-processing), (2) 3D Image deformation (image warping) (3) Hole filling (hole filling). Stereo image synthesizing method of this study was to explore effective and new technology made a hole to fill. Balcerak et al. [1] presents a red component of depth map generation and hole filling method in monitoring system of 3D effects. Its binary depth map is based on target recognition and tracking, and use the red component is mirrored and filter (blocking) hole filling method. Camplani and Salgado [2] the proposed method is based on a joint bi-directional filtering framework, including space and time information, with their repeated use of adjacent pixels depth of joint bilateral filter to recover their lost values. Chen et al. [3] within a depth of to the edge of the filter and put effective depth image of the drawing. It can effectively solve DIBR hole in the system fill problem. Cheng et al. [4] present three view image generation technologies to improve results. First is to use bilateral filters in depth on the image, it may help the continuous change of the depth and smoothness similar adjacent color depth, so noise can inhibit the image distortion. Secondly, in variants a new observation point on the image, depth to depth strata in order to fill the background of the shadowed area on the image in order to maintain depth structure. For color images, image drawing based on the depth-oriented examples, combines the structural strength of gradient colors, image to maintain schema in the reply area. Last triangular filter combined with the spatial position, color, intensity and depth of information to determine weight, enhances image synthesis of results. Choi et al. [5] they recommended depth estimation and image restoration of high quality 3D video hole filling

method. First of all, they draw images and the depth maps. Then, there was a hole estimated in the depth map. Auxiliary images based on sparseness and depth of holes to fill to fix these holes. Cigla and Alatan [6] using reliable horizontal and vertical orientation of the texture transform from visible pixel to fill holes. They use similar to that of neighboring pixels and depth of color continuous weighted sum (SWS), the adaptive weighting and connection aggregation. Dong et al. [7] analyze on DIBR virtual view of the drawing process, two views presented to modify the parallax method for generating new views and combines the image patch to improve. Parallax is determined by the depth of the above information to infer, based on disparity gradient correction, thus reducing previous 3-D image deformation produced a new hole in the view area. Doria and Radke [8] combined the images from the supplement the basic concept of patch and the gradient field image editing, both in texture and structure of two species of fill light (LiDAR) to scan images. Once the depth gradient, they use image reconstruction techniques, to obtain the final 3D scene structure. Mokhtar and Pramod [9] used two new 3 by 3 structural elements can speed up the modified hole-filling algorithm. Hsiao et al. in [10] a new algorithm is presented, along with variants and holes to fill, making calculated overall significant reduction in the delay. Two ways to upgrade around the edges of parallax a horizontally mirrored is used to reduce the synthesis of virtual images generated by the visual flaws. Hung and Siu [11] propose a depth aided nonlocal-means algorithm, using information from the current frame and another frame figure to fill the hole in the synthesis of video. Hwang et al. [12] they recommended image based image segmentation method of patching holes to fill. Their holes filled can not only show the high quality of the results, can also be applied on the commonly used methods for real-time applications. Jung and Ho [13] make a hole filling algorithm using spatial domain of adjacent frames with 3D video synthesis of color and depth of deformation and eye hole in the depth of the linear interpolation. As the interpolated depth value, search in the time domain corresponds to the colors and the textures used to fill in the hole in the color image. Mao et al. [14] provide the first observation hole and fill the hole of expansion problems. First, proposed a depth to convert pixel histogram to identify missing or incorrect expansion arising out of holes. Then, put forward two different computational complexity of technology to fill the expanding holes (1) linear interpolation, and (2) based on graphical interpolation with sparseness of priority. An adaptive method of image distortion to improve the general method of deformation was presented by Plath et al. [15]. The algorithm is used to smooth a typical depth map, in order to reduce the potential for optimization problems. To prevent deformation of the hole must be found. This depth map divided into blocks of uniform depth, use the four-element tree and perform adaptive grid to achieve optimal results. Solh and AlRegib [16] proposed removing shaded areas DIBR with two new methods. Respectively, these two approaches are hierarchical holes filled (HHF) and the holes filled with Adaptive depth hierarchy to eliminate the depth map by any smoothing or filtering needs repair. Both

technologies use a pyramid method to estimate, from 3D deformation of hole in the estimation of low resolution pixels in the image. The estimated the low resolution images to virtual zeros deleted, together with deformation of the Gaussian filter. With adaptive depth of HHF using the depth information result in a higher resolution to previously shaded areas of the drawing. Wang et al. [17] provide an asymmetric edge adaptive filter (AEAF) partially addressed two 3DTV Challenges, namely depth to generate and fill the hole. Unlike other similar treatment is, AEAF operation can be achieved when edge effects of pretreatment of correction and depth charts. Wang et al. [18] put forward based on projection onto convex sets a new hole to fill. They will observe the way pixels projected into the hole in the frame of reference of the frame formed by the convex set to fill the position of holes by it. Xu et al. [19] show the filtering shaded areas of the holes by Kinect photography method of the depth map. This method is used by traditional Kinect camera and depth maps provided by the original RGB image. They used the original image capturing moving objects and background differences, and then filled the main application area for 4-neighborhood pixels interpolated to holes area. Xu et al. [20] proposed a deep secondary image restoration method based on the model to restore a large shaded area. It includes two processing, deformation depth maps to fill and color image to fill. Because depth map can be considered without texture, gray-scale images, it is easy to fill. Shaded areas of the color images are based on the associated filling depth map information to predict. Background area with the texture has a high priority in filling and shaded areas by model-based image restoration with its background texture fill. First of all, Yang et al. [21] the proposed method to 8 connecting depth as a benchmark the hole number. For each hole in the depth of a number, use deep holes neighborhood pixel depth distributions for filling deep holes. Then, use the improved cross-bidirectional filter to fill the holes. Simply use the depth distribution of neighboring pixels, this method to improve the depth map, and reduced depth due to incorrect color filled.

In this paper, we give the general procedure of stereoscopic image generation and a hole filling technique. Next section we show some basic previous works that we are going to apply. The third section we demonstrate the computer experiment results. The conclusions and comments are given in the final section.

2 Proposed method

Currently it often used in 2D to 3D converted image synthesis techniques for depth imaging of mapmaking. Depth image drawing method is based on stereo system for objects derived from stereo image synthesizing method that adds up, the main process consisted of three parts: (1) pre-processing, (2) 3D image warping (3) hole filling. Pre-processing uses mainly the depth map by smooth filter after processing, noise reduction, and decides to zero plane (left and right eye images of parallax for the 0 position) location, General views of zero plane Multi-DVD set for the depth value is 128 or 255

points. Fig. 1, as a parallel set of diagrams of the photo camera and stereo image synthesis, C_l and C_r severally for the optical center of the camera in the right eye and left eye (optical Center), C_c the optical center of the camera shot for us.

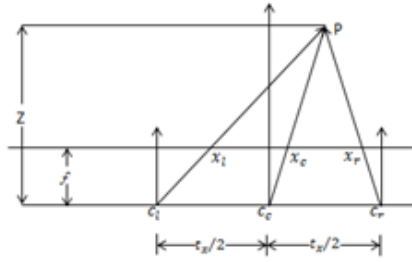


Fig. 1 a parallel set of diagrams of the photo camera and stereo image synthesis

C_c camera image as the middle of the left and right eye image, and after the calculated horizontal displacement to the left and to the right, it will be able to synthesize images shot left and right cameras, as shown in the formula is as follows:

$$x_l = x_e + \frac{t_x f}{2 Z} \tag{1}$$

$$x_r = x_e + \frac{t_x f}{2 Z} \tag{2}$$

where x_l and x_r is to synthesize the pixels located in the left and right eye images; x_c coordinates from the image in the middle, center image location, f is focal length, experiment is set to a constant; t_x is the baseline length; Z is the p location depth value, contrary to depth grayscale values (that is, Z smaller, greater depth grayscale value). By the formula, deep closer, larger the parallax, the coordinates on the image left image shifts to the right and vice versa.

Regardless of the advantages, it has a serious problem. When we synthesize a virtual view with single color and depth image, occlusion regions appear. The occlusion is a visible region created due to different view point position as shown in Fig. 2. Since the occlusion regions have no information in input data, they are shown like holes on the synthesized images.

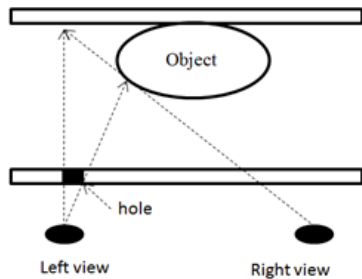


Fig. 2 the hole appearance in image synthesis

2.1 Review of mathematical morphology

1. The erosion of the binary image A by the structuring element B is defined by:

$$A \ominus B = \{z \in E | B_z \subseteq A\}, \tag{3}$$

where B_z is the translation of B by the vector z ,

$$\text{i.e., } B_z = \{b + z | b \in B\}, \forall z \in E. \tag{4}$$

2. The dilation of A by the structuring element B is defined by:

$$A \oplus B = \bigcup_{b \in B} A_b \tag{5}$$

The dilation is commutative, also given by :

$$A \oplus B = B \oplus A = \bigcup_{\alpha \in A} B_\alpha \tag{6}$$

3. The opening of A by B is obtained by the erosion of A by B , followed by dilation of the resulting image by B :

$$A \circ B = (A \ominus B) \oplus B. \tag{7}$$

4. The closing of A by B is obtained by the dilation of A by B , followed by erosion of the resulting structure by B :

$$A \bullet B = (A \oplus B) \ominus B. \tag{8}$$

2.2 Review of bilateral filter:

The bilateral filter is defined as

$$I^{filtered}(x) = \frac{1}{W_p} \sum_{x_i \in \Omega} I(x_i) f_r(\|I(x_i) - I(x)\|) g_s(\|x_i - x\|) \tag{9}$$

where the normalization term

$$W_p = \sum_{x_i \in \Omega} f_r(\|I(x_i) - I(x)\|) g_s(\|x_i - x\|) \tag{10}$$

ensures that the filter preserves image energy and

- $I^{filtered}$ is the filtered image;
- I is the original input image to be filtered;
- x are the coordinates of the current pixel to be filtered;
- Ω is the window centered in x ;
- f_r is the range kernel for smoothing differences in intensities. This function can be a Gaussian function;
- g_s is the spatial kernel for smoothing differences in coordinates. This function can be a Gaussian function;

2.3 Proposed method

In this method there are two of the most important parts is the acquisition of depth information, another is the hole

filling after image warping. For acquisition of depth information edge information can be used to convert color images to grayscale depth map. Another problem for the depth maps for filling of holes you can do preprocessing. Preprocessing for depth map is doing image smoothing. It can reduce the number and scope of holes. The flow chart of our proposed as following:

Step 1: Depth image drawing method: 2D images into 3D images using depth image drawing method, as shown in Figure 3.

Step 2: Use depth information classification, the distinction between foreground and background: Mark the foreground area is to search for matching blocks can be judged, and morphological characteristics with a depth map of the closed holes filled up, so the blocks will be marked future errors can be avoided by taking to fill.

Step 3: Detecting Hole type: it makes different situation for different processing; by single image synthesis produces more virtual images, this process caused many images of broken situation, roughly divided into single points of broken hole, small range cracks, and big regional broken hole, its causes for rounded of calculation errors value, as referred to around eye calculation out of formula, and depth changes not obviously regional and prospects and background junction, depth dramatic changes.

Step 4: For a small break area get holes and cracks in the small, mean approach, breaking scenario for minor uses a simple way to do it, can effectively reduce the operation time, while maintaining the image quality.

Step 5: Large broken area filled: For a hole in the image to fill the order, select a gradient higher the value, the higher the priority, the aim is to give priority to the image partially filled with texture information to complete, avoiding vertical or horizontal lines break, imaging of the unnatural. Observation on imaging characteristics of holes, found forming holes due to image after Warping caused by horizontal displacement to the left or to the right, so the best way to fill is the holes around the blocks to find the closest replication similar to determine whether the sum of squared differences (SSD) as a judge and representatives of the smaller the number, the more similar. We use morphological holes filled at this time.

The procedure of our proposed method is shown in Figure 3.

3 Experiments Results

The experiment of the proposed algorithm, we apply it to one-view and one-depth video of the benchmark image "Ballet" shown in Fig. 4.

Fig. 5 shows the labeled of foreground of depth map after lateral filtering. We synthesize the single color image

using 3D warping with the camera parameters. Fig. 6(a) is the left view of ballet and Fig. 6(b) is the right view of ballet after warping. Fig. 7 and Fig. 8 show the final results.

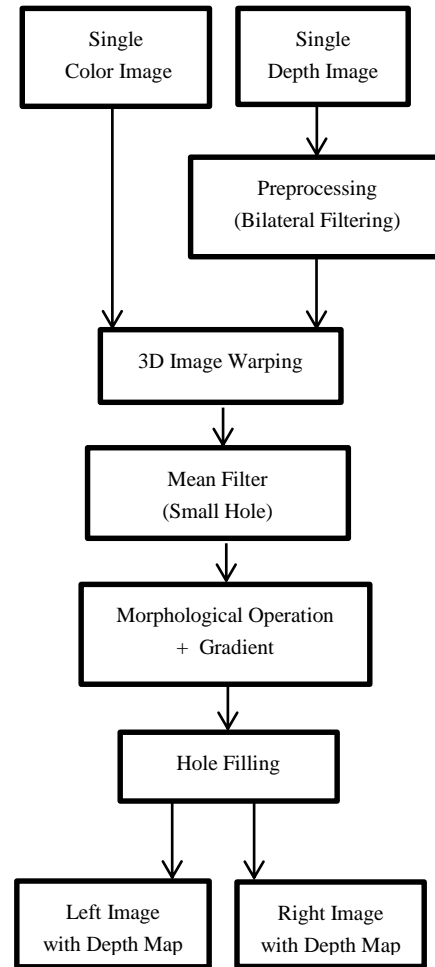


Fig. 3 the flow chart of proposed method

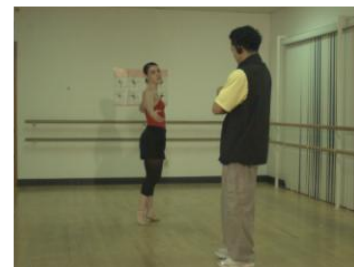


Fig.4 single color image Ballet

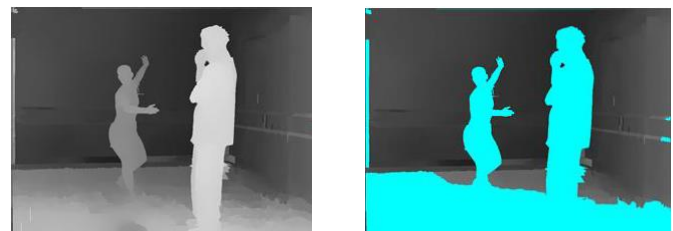


Fig. 5 (a) depth map after lateral filtering (b) labeled of foreground



Fig. 6(a) left view image after warping (b) right view image after warping



Fig. 7(a) left view image after hole filling (b) right view image after warping



Fig. 8(a) left view depth after hole filling (b) right view depth after warping

4 Conclusions

Proposed method in this paper, we select the right way according to various characteristics of holes to fill, and use the morphology, texture information area for the highest priority, and also refer to the depth of information, avoid to use the foreground information to fill the hole in the background error conditions occur. Through experimental results demonstrate that this method effective in the benchmark scenes and complete hole repair images, which allows us to enjoy a more natural and proper 3D images of the future 2D convert to 3D technologies become more widespread.

Acknowledgment

This paper is supported by HCU-102-B1-08.

5 References

[1] Balcerek, J., Konieczka, A., Dabrowski, A., & Marciniak, T. "Binary Depth Map Generation and Color Component Hole Filling for 3D Effects in Monitoring Systems". In *Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA)*, 2011 (pp. 1-6).IEEE, Sep 2011.

[2] Camplani, M., & Salgado, L. "Efficient Spatio-Temporal Hole Filling Strategy for Kinect Depth Maps". In *IS&T/SPIE Electronic Imaging* (pp. 82900E-82900E). International Society for Optics and Photonics, Feb 2012.

[3] Chen, W. Y., Chang, Y. L., Lin, S. F., Ding, L. F., & Chen, L. G. "Efficient Depth Image Based Rendering With Edge Dependent Depth Filter and Interpolation". In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on* (pp. 1314-1317). IEEE, Jul 2005.

[4] Cheng, C. M., Lin, S. J., Lai, S. H., & Yang, J. C. "Improved Novel View Synthesis From Depth Image with Large Baseline". In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (pp. 1-4). IEEE, Dec 2008.

[5] Choi, J., Choe, Y., & Kim, Y. G. "Sparsity Based Depth Estimation And Hole-Filling Algorithm for 2D to 3D Video Conversion". In *Signals and Electronic Systems (ICSSES), 2012 International Conference on* (pp. 1-4).IEEE, Sep 2012.

[6] Cigla, C., & Alatan, A. A. "An Efficient Hole Filling for Depth Image Based Rendering". In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on* (pp. 1-6). IEEE, Jul 2013.

[7] Dong, H., Jianfei, S., & Ping, X. "Improvement of Virtual View Rendering Based on Depth Image". In *Image and Graphics (ICIG), 2011 Sixth International Conference on* (pp. 254-257). IEEE, Aug 2011.

[8] Doria, D., & Radke, R. J. "Filling Large Holes in LiDAR Data by Inpainting Depth Gradients". In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on* (pp. 65-72). IEEE, Jun 2012.

[9] Hasan, M. M. & Mishra, P. K. "Improving Morphology Operation for 2D Hole Filling Algorithm". *International Journal of Image Processing (IJIP)*, Volume (6) : Issue (1) : 2012 (pp. 1-12), 2012.

[10] Hsiao, S. F., Cheng, J. W., Wang, W. L., & Yeh, G. F. "Low Latency Design of Depth-Image-Based Rendering Using Hybrid Warping and Hole-Filling". In *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on* (pp. 608-611). IEEE, May 2012.

[11] Hung, K. W., & Siu, W. C. "Depth-Assisted Nonlocal Means Hole Filling for Novel View Synthesis". In *Image Processing (ICIP), 2012 19th IEEE International Conference on* (pp. 2737-2740). IEEE, Sep 2012.

[12] Hwang, J., Lee, K., Kim, J., & Lee, S. "A Novel Hole Filling Method Using Image Segmentation-Based Image In-Painting". In *Consumer Electronics (ICCE), 2013 IEEE International Conference on* (pp. 470-471). IEEE, Jan 2013.

- [13] Jung, J. I., & Ho, Y. S. "A Hole Filling Technique in The Temporal Domain for Stereoscopic Video Generation". In Proceedings of 2011 APSIPA Annual Summit and Conference, Xian, China, Oct 2011.
- [14] Mao, Y., Cheung, G., Ortega, A., & Ji, Y. "Expansion Hole Filling in Depth-Image-Based Rendering Using Graph-Based Interpolation". In Acoustics, Speech and Signal Processing, IEEE International Conference on, Vancouver, Canada, May 2013.
- [15] Plath, N., Knorr, S., Goldmann, L., & Sikora, T. "Adaptive Image Warping for Hole Prevention in 3D View Synthesis". IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 2013.
- [16] Solh, M., & AlRegib, G. "Hierarchical Hole-Filling For Depth-Based View Synthesis in FTV and 3D Video". Selected Topics in Signal Processing, IEEE Journal of, 6(5), 495-504, 2012.
- [17] Wang, L. H., Huang, X. J., Xi, M., Li, D. X., & Zhang, M. "An Asymmetric Edge Adaptive Filter for Depth Generation and Hole Filling in 3DTV". Broadcasting, IEEE Transactions on, 56(3), 425-431, 2010.
- [18] Wang, W., Yang, Y., & Liang, Q. "A Novel Hole Filling Method Based on Projection onto Convex Set in DIBR". In 3rd International Conference on Multimedia Technology (ICMT-13). Atlantis Press, Nov 2013.
- [19] Xu, K., Zhou, J., & Wang, Z. "A Method of Hole-Filling for The Depth Map Generated by Kinect with Moving Objects Detection". In Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on (pp. 1-5). IEEE, Jun 2012.
- [20] Xu, X., Po, L. M., Cheung, C. H., Feng, L., Ng, K. H., & Cheung, K. W. "Depth-Aided Exemplar-Based Hole Filling for DIBR View Synthesis". In Circuits and Systems (ISCAS), 2013 IEEE International Symposium on (pp. 2840-2843). IEEE, May 2013.
- [21] Yang, N. E., Kim, Y. G., & Park, R. H. "Depth Hole Filling Using The Depth Distribution of Neighboring Regions of Depth Holes in The Kinect Sensor". In Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on (pp. 658-661). IEEE, Aug 2012.

Invariant Feature Extraction using 3D Silhouette Modeling

Jaehwan Lee¹, Sook Yoon², and Dong Sun Park³

¹Department of Electronic Engineering, Chonbuk National University, Korea

²Department of Multimedia Engineering, Mokpo National University, Korea

³IT Convergence Research Center, Chonbuk National University,

Abstract - One of the major challenging tasks in object recognition results from the great change of object appearance in the process of perspectively projecting objects from 3-dimensional space onto 2-dimensional image plane with different viewpoints. In this paper, we proposed a method to extract features invariant to limited movements of objects by constructing a 3-D model using silhouettes of objects from images with multiple viewpoints. We investigated several renowned invariant features to find the most appropriate one for the proposed method, including SIFT[5], SURF[6], ORB[7], BRISK[8]. The simulation results shows that all the invariant features tested work well and the SURF performs best in terms of matching accuracy.

Keywords: Invariant Feature, Shape From Silhouette, Intelligence Surveillance System

1 Introduction

Accurate recognition of 3-dimensional objects in 2-dimensional images is the most crucial and difficult task in image understanding. There are many possible factors making the recognition task challenging such as information loss from perspective transformation, illumination effects and various appearance of non-rigid body objects[1]. Especially, movements of non-rigid objects in the 3-D space may significantly change images of objects so that matching models in the database with input object images may experience a large difference.

There have been many techniques to resolve the difficulties by using color information, face recognition, part-based recognition, video based gait recognition[2][3][4], etc. Popular image-based local feature description methods such as SIFT[5], SURF[6], ORB[7], BRISK[8] are used to deal with the movements of objects by designing the features invariant to the appearance changes. These feature description methods may work well for a given situation, however, the matching accuracy may need to be improved for the case of recognizing very flexible objects with various viewpoints.

In this paper, We proposed an invariant feature extraction method using a 3-D modelling based on silhouettes of objects from multiple images with different viewpoints. The method firstly construct 3-D models with shape from silhouette approach and then use these models to extract invariant features applicable for any viewpoint. To determine the best feature description method for the proposed method, we also investigate several state-of-the-art feature description methods including the four image-based local feature description methods mentioned above.

2 Proposed Feature Extraction Method

The overall block diagram of the proposed extraction method is depicted in Fig. 1. It consists of a feature extraction block and a test phase block. The first clock is to construct a 3-D model with multiple images and to extract features after projecting the constructed model onto a 2-D plane according to the angle obtained in the pose detection step. This feature extraction method can generate features for any viewpoint changes that can be used to compare to the actual features from test images.

We reconstructed 3D models using the Shape From Silhouette (SFS), described in the Ref. 15, which requires relatively fewer images than other 3-D modelling techniques. Shape From Silhouette is a shape reconstruction method which constructs a 3D shape estimate of an object using silhouette images of the object[16]. In this step, 3D model is trained using multiple images. A set of reconstructed 3-D models can be stored in a database each representing an object at the training phase. These models can be projected onto any 2-D image plane with a specific viewpoint and to be used to extract local features using one of the existing popular methods such as ORB, BRISK, SIFT and SURF shown at the second and third steps. The two steps are later used to verify the existence of objects for test images with additional information from the test stage.

At the test stage, new test images are presented to the system. A test image is firstly used to segment out object regions and then extract features for the regions. The extracted features from the current input image are compared with those from the 3-D models with a set of viewpoints for

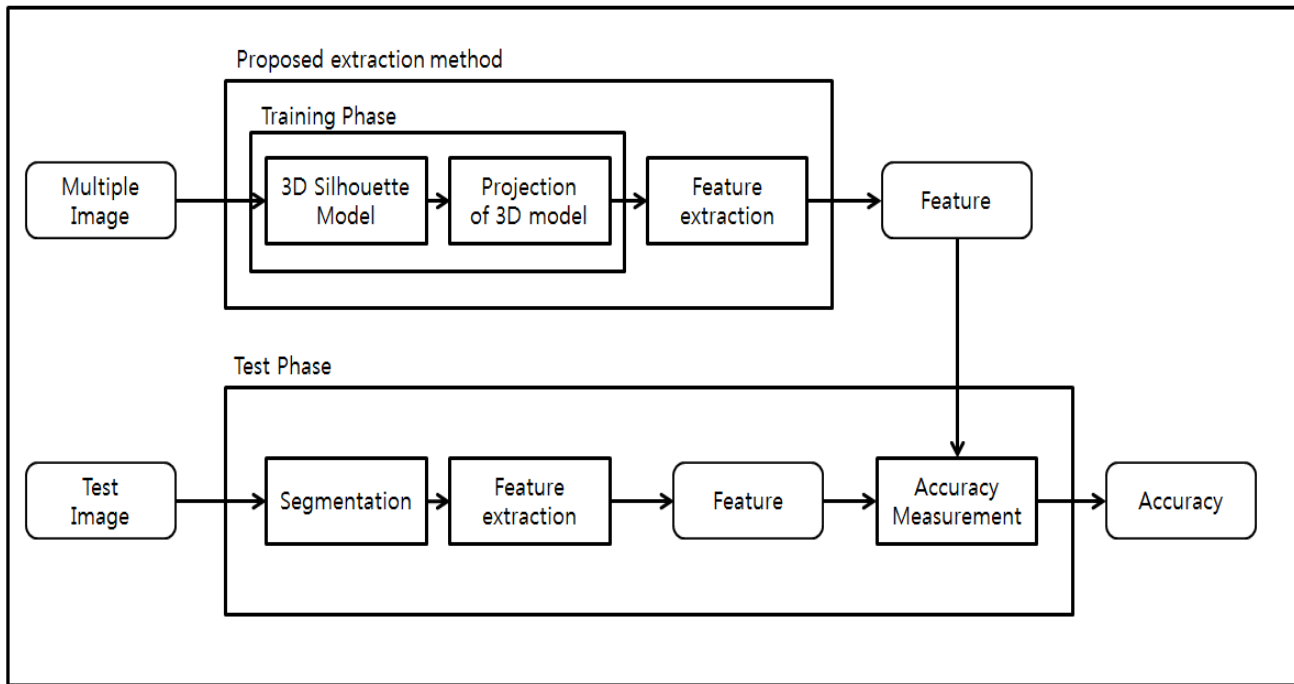


Figure 1. Overall Block Diagram

matching. If a maximum matching score above a certain threshold value, we accept the input image containing the specific object with a viewpoint. There can be a series of comparison to find the best possible matching.

In this paper, we focused on selecting the most appropriate feature description method which shows the highest matching accuracy. We used four feature description methods: ORB, BRISK, SIFT and SURF. For this purpose, we use the input data set with known viewpoints representing the target objects and compare these objects to those objects reconstructed from the 3-D models. Each feature description method used in this paper generates two sets of feature points for a target object and a reconstructed object. Then each feature point in a target set searches for a feature point in another set for matching. Matching of a point is defined as true if the relative distance between two points is less than a predefined threshold value. The matching accuracy between the two sets of feature points is then determined as in Eq. 1.

$$\begin{aligned}
 TP &= \text{Number of true matching feature points} \\
 FP &= \text{Number of false matching feature points} \quad (1)
 \end{aligned}$$

$$\text{Matching Accuracy} = \frac{TP}{TP + FP}$$

Fig. 2 shows an example of true and false matching. The objects on the left and right are from input reference image and reconstructed image with reduced size, respectively. In the figure, a false matching and a true matching between feature points are shown as red and blue line segments.

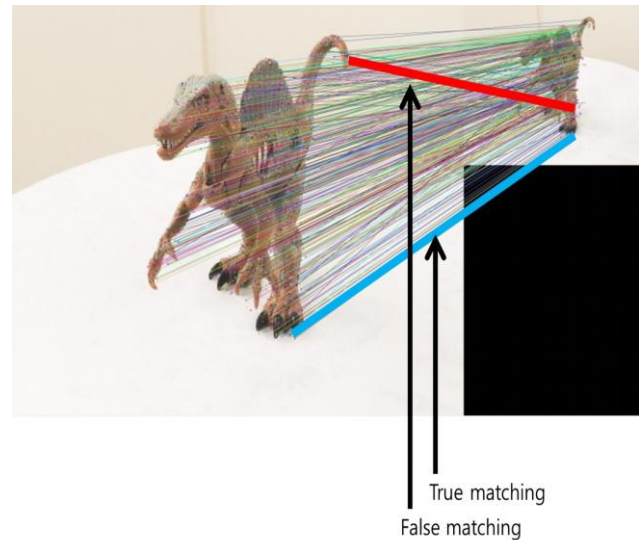


Figure 2. True and False matching

3 Experiments and Discussions

Two data sets, Visual Geometry Group data set[13] and Yasutaka Furukawa and Jean Ponce data set[14], are used for the experiment. The Visual Geometry Group data set contains 36 720x576 images for an object all with different viewpoints. The Yasutaka Furukawa and Jean Ponce data set also contains 24 200x1500 images for another object. Fig 3 shows two example images from each data set. We used the shape from silhouette(SFS) functions introduced by Lore Shure[11] to perform the 3D modeling and the projection of the constructed 3D model. We used feature description methods

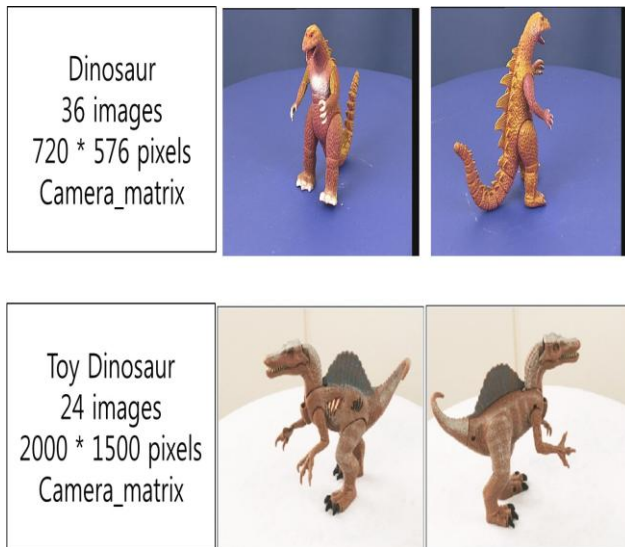


Figure 3. Used image for reconstructing 3D model
 (up)Visual Geiometry Group data set (down)Yasutaka
 Furukawa and Jean Ponce data set

implemented in opencv library to extract local features for ORB, BRISK, SIFT and SURF.

3.1 3D Silhouette Modeling

A 3-D model of an object is reconstructed with different number of images, using the SFS. For this experiment, we tested with 4, 8 or 36 images for the reconstruction. The angles between two images become 90° , 45° , and 10° for 4, 8, and 36 training images, respectively.

Fig. 4 shows the 3D modeling results of the Visual Geometry Group data set. Three 3D models are reconstructed first with three different number of images and the models are projected for two different viewpoints. For the original images as targets with two different viewpoints, shown in Fig. 4a, the projected images from three 3-D models shown in Fig.4 (b)-(d). As we can expect, the more images with different viewpoints are used, the better the reconstructed image quality is. Fig. 5 shows the 3-D modelling results the Yasutaka Furukawa and Jean Ponce data set.

3.2 Matching Accuracy of Feature Extraction Methods

Before testing the matching accuracy of the invariant features from 3-D modelling, we executed a simple effectiveness and accuracy test to each the feature extraction method. To know the performance of each method, a reference image is modified with a gaussian smoothing and a resizing operations and the matching between the original and the modified versions are performed. Fig. 6 shows the images used for this experiment. Fig. 6a-c shows the 256×256 original image, and the modified images with Gaussian smoothing and the resizing to half size.

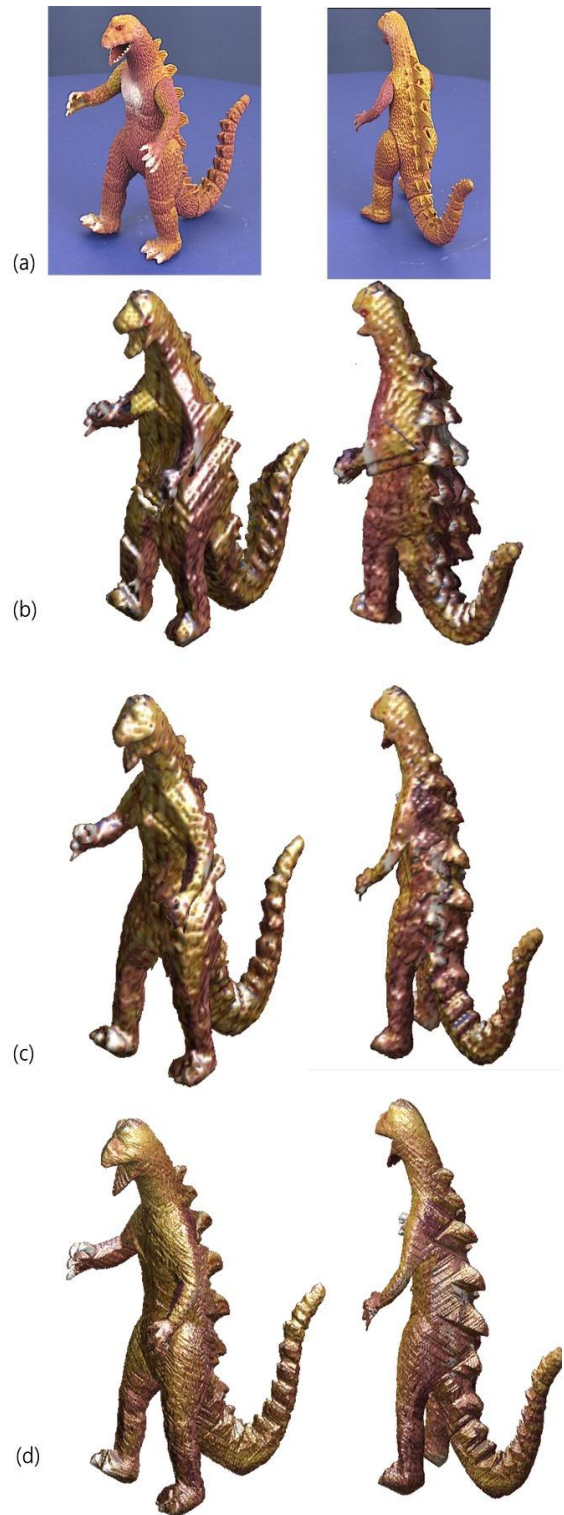


Figure 4. Reconstructed 3D model using Visual
 Geometry Group data set



Figure 5. Reconstructed 3D model using Yasutaka Furukawad and Jean Ponce data set

Table 1 shows the accuracy measurement results. In the experiment, a matching is defined as true if the distance between locations of two feature points is less than 10 pixels. As we can see in the table, BRISK extracts the less number of feature points and the accuracy is very low. For the case of SURF, we assume that it produces enough number of feature points but the accuracy is not high enough. The number of extracted features of ORB, SURF and SIFT are abundant, the accuracy is relatively high. Thus we assume that ORB and SIFT are good features to use for these types of modifications. Especially, the SURF is very robust to smoothing operation and the SIFT is robust to resizing operation.

Fig. 7 shows an example of feature matching between a projected image from its 3D model reconstructed with 8 images and the corresponding target image. Experimental results show rather lower accuracy than the simple image example. In case of comparison between the actual image and the projected image from a 3D model, SIFT has best accuracy. In this experiment, we measure the accuracy between images from a 3D model with more viewpoints and images from a 3D model with a less number of viewpoints. The reason of this comparison is to produce target images with more details than

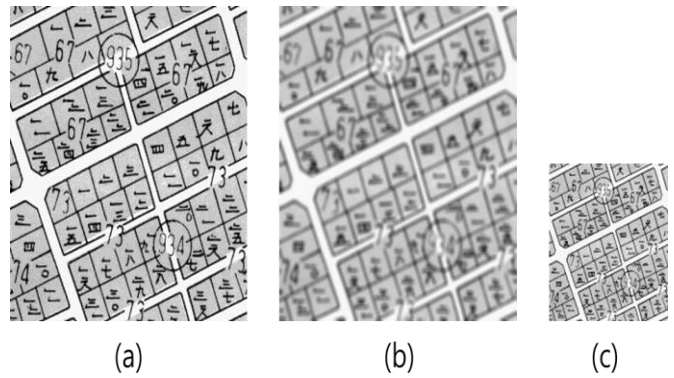


Figure 6. Images for simple matching accuracy

	Smoothing		Resizing	
	Number of target's features	Accuracy (%)	Number of target's features	Accuracy (%)
ORB	460	100	240	82
BRISK	173	9	58	0.6
SIFT	1197	95	341	100
SURF	569	73	163	47

Table 1. Matching accuracy for simple 2-D image

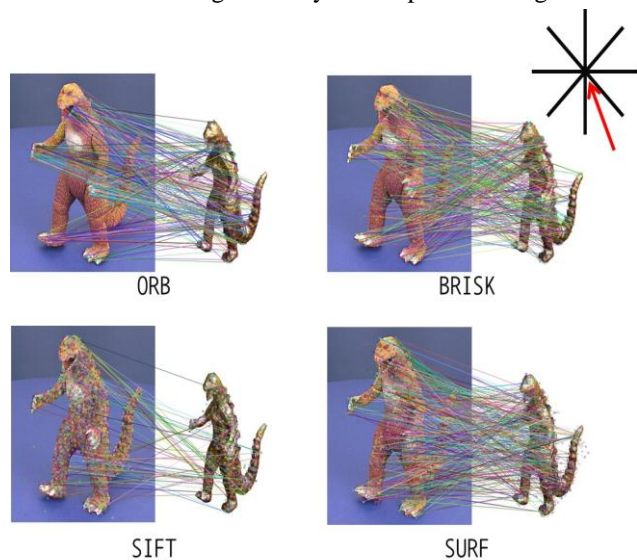


Figure 7. Feature Matching example between target Image and a projected image from 3D model

test images. In case of comparison between images from a 3D model and another 3D model, SURF has best accuracy, over ORB and SIFT.

4 Conclusion

Identifying an 3-dimensional object appeared in different angles is a very challenging task in computer vision, even if

			Accuracy (%)			
	Reference	Target	ORB	BRISK	SIFT	SURF
VGG	Real	4 Images	0.73	2.16	6.61	2.32
	Real	8 Images	0.34	1.46	4.29	3.68
	36 Images	4 Images	18.75	18.7	23.08	24.2515
	36 Images	8 Images	34.74	17.62	33.33	38.157
YF&JP	Real	4 Images	3.94	6.34	5.56	5.60
	8 Images	4 Images	38.30	16.42	16.67	13.56

Table 2. The accuracy of feature about reconstructed model

we exclude external factors making the recognition even worse. In this paper, we used the shape from silhouette technique to reconstruct 3-D models of objects with multiple images. The reconstructed 3-D model is used to produce a 2-D projected image with a specific viewpoint for comparison using renowned feature description methods, including ORB, BRISK, SIFT and SURF. The reconstructed 3D models from multiple images contains more details as the more training images are used. When a better reconstructed model is used for testing, the matching accuracy becomes higher. Although there are some positive evidences for automatic generation of invariant features using 3-D modelling, generally speaking, matching performance of feature points are not good enough for the current set of feature extraction methods and different types of features should be developed for this purpose.

We will further search for better features and 3-D models to automatically generate invariant features using 3D models.

5 Acknowledgement

This work was supported by the Brain Korea 21 PLUS Project, National Research Foundation of Korea and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2013R1A1A2013778).

6 References

- [1] S. Fleck and W. Straber, "Privacy sensitive surveillance for assisted living a smart camera approach," Handbook of Ambient Intelligence and Smart Environments, Springer, pp.985-1014, 2010
- [2] Amit A. Kale, Aravind Sundaresan, A. N. Rajagopalan, Naresh P. Cuntoor, Amit K. Roy Chowdhury, Volker Kruger, and Rama Chellappa. "Identification of humans using gait", IEEE Transactions on Image Processing, 13(9):1163-1173, September 2004.
- [3] Alper Yilmaz, Omar Javed, and Mubarak Shah, "Object Tracking: A Survey", ACM Computer Surveys, Vol.38, No.4, Article 13, Publication date: December 2006
- [4] Laurenz Wiskott, Jean-Marc Fellous, Norbert Kruger, and Christoph von der Malsburg, "Face recognition by elastic Bunch graph matching", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 19, NO. 7, JULY 1997
- [5] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60, 2, pp. 91-110, 2004
- [6] H. Bay, T. Tuytelaars, and L. Van Gool. "SURF: Speeded up robust features.", Computer Vision–ECCV 2006, pages 404–417, 2006
- [7] Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary R. Bradski, "ORB: An efficient alternative to SIFT or SURF". ICCV 2011: 2564-2571.
- [8] Stefan Leutenegger, Margarita Chli, Roland Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," iccv, pp.2548-2555, 2011 International Conference on Computer Vision, 2011
- [9] Pierre Moreels and Pietro Perona, "Evaluation of Features Detectors and Descriptors based on 3D objects", ICCV2005, Vol.1, pp.800-807, 2005
- [10] Powers, David M W, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies, Vol.2, Issue.1, pp.37–63, 1970
- [11] Loren Shure, "Carving a Dinosaur", <http://blogs.mathworks.com/loren/2009/12/16/carving-a-dinosaur/>, [Access: 2014.04.19]

[12] OpenCV User Site, <http://opencv.org> [Access: 2014.05.19]

[13] Visual Geometry Group, "Dino data", Department of Science, University of Oxford, <http://www.robots.ox.ac.uk/~vgg/data1.html> [Access:2014.04.19]

[14] Yasutaka Furukawa and Jean Ponce, "3D Photography Dataset", Beckman Institute and Department of Computer Science University of Illinois at Urbana-Champaign

[15] Gloria Haro, "Shape from Silhouette Consensus", *Pattern Recognition*, Vol. 45, No. 9, pp. 3231-3244, 2012

[16] Kong-man (German) Cheung, Simon Baker and Takeo Kanade, "Shape-from-Silhouette Across Time - Part I: Theory and Algorithms", *International Journal of Computer Vision*, Vol. 63, pp. 225-245,

SESSION

PATTERN RECOGNITION, IMAGE RETRIEVAL METHODS, IMAGE FEATURE EXTRACTION AND FEATURE MAPPING, IMAGE SEGMENTATION, EDGE DETECTION

Chair(s)

TBA

Automatic Segmentation of Coronal Mass Ejections from STEREO white-light coronagraph images

V. Kirnosov¹, L.-C. Chang¹, and A. Pulkkinen²

¹Electrical Engineering & Computer Science Dept., The Catholic University of America, Washington, D.C., USA

²NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

Abstract - *In this paper we propose a new technique to segment Coronal Mass Ejection (CME) using STEREO SECCHI COR2 beacon images. The beacon data are usually used for space weather forecasting; the data were highly compressed for faster transmission and therefore have lower image resolution and poor quality. Moreover, according to our analysis, the data suffer from intensity level instability that added challenges for automatic near real-time CME detection. The method proposed in this paper overcomes this problem by classifying images into two classes. Segmentation is done using image subtraction technique, followed by morphological erosion and histogram equalization. The method was validated using fifteen CME events with series of FITS images from STEREO A/B spacecraft. The result shows that proposed method is effective for noise reduction and CME segmentation. The segmented CME is useful for further automatic CME leading edge detection and can improve the accuracy of propagation parameters derivation.*

Keywords: STEREO; Data Analysis; Coronal Mass Ejection; Automatic Image Segmentation; Space Weather

1 Introduction

CMEs are large-scale expulsions of plasma and magnetic field from the solar atmosphere and have been recognized as primary drivers of interplanetary disturbances [1]. Until the early years of this century, images of CMEs had been made near the Sun primarily by coronagraphs on board spacecraft. Coronagraphs view for the outward flow of density structures emanating from the Sun by observing Thomson-scattered sunlight from the free electrons in corona and heliospheric plasma [2]. The National Aeronautics and Space Administration (NASA), the European Space Agency (ESA), and other nations' space programs have launched several missions specifically to observe these solar phenomena [3].

In late 1995, joint NASA/ESA SOHO mission was launched, and two of its three LASCO coronagraphs still operate today. LASCO was joined by NASA twin Solar Terrestrial Relations Observatory (STEREO) [4] A/B coronagraphs in 2006 [2]. STEREO uses two spacecraft with identical instrumentation consisting of a series of coronagraphs and heliospheric imagers to track CMEs

simultaneously from two vantage points. One spacecraft is leading the Earth (STEREO A) and the other is trailing (STEREO B) [5].

The key objective of these coronagraphs was to understand the initiation and propagation of CMEs, and their effect on the near-Earth environment. To accomplish these goals, it has been acknowledged that the three-dimensional nature of the corona must be observed, i.e., multiple spacecraft with multiple viewing angles at coronal and heliospheric phenomena were required [4].

STEREO has two separate telemetry streams coming down from each spacecraft, the space weather beacon telemetry, and the science recorder playback telemetry. The beacon telemetry contains the most recent data and images, and is transmitted 24 hours per day. The beacon telemetry rate is very low, the images need to be compressed by large factors, and are thus of much lower quality than the actual science data that reach the STEREO Science Center with the delay of several days [6]. It is the policy on the STEREO project that all data be available within a short period of time [4].

For near real-time CME detection it is crucial to get and analyze recent coronagraph data as soon as possible, so beacon images are the most suitable for this purpose. Beacon data are available in 2 formats: FITS and JPEG image files. JPEG beacon files are 8-bit RGB color images (three color channels, each channel has 256 possible intensity levels) and FITS files are 16-bit grayscale images (single channel, 65536 possible intensity levels).

Detection of CMEs has traditionally been addressed by visually checking coronagraph data for outward moving features [1]. The visual detection of CMEs in the flood of incoming new data is a labor intensive task [7], so these "manual" catalogs [8] have in recent times been augmented by additional catalogs [9 - 11] of CMEs detected by automatic methods [2]. Methods used in these catalogs detect CMEs from SOHO LASCO white-light coronagraph images. In addition to that, CACTus [10] and SEEDS [11] catalogs are also capable to process STEREO COR2 data and provide propagation parameters that describe events. As an advanced feature, the method used in SEEDS catalog allows to detect

front edge of the CMEs. This method was added to the SEEDS very recently and it has not yet been rigorously tested [12].

To the best of our knowledge, there is a shortage of publications and methods for CME front edge detection using STEREO white-light coronagraph images, which prompted us to address this issue. The finding in this paper is the fundamental knowledge for the CME front edge detection. Our ultimate goal is to use the detected edges from images provided by two STEREO spacecraft and perform automated triangulation of the CME parameters.

In this paper, we inspect and investigate a huge amount of STEREO COR2 images, and report our findings related to image quality. We also propose a novel technique for efficient background removal and CME segmentation suitable for further CME leading edge detection and extraction of parameters. We validate our approach using the data set of fifteen CME events with series of COR2 images from STEREO A/B spacecraft.

2 Data Analysis

During thorough visual analysis of STEREO A/B COR2 beacon files, we noticed periodic intensity level variations between the images. This issue is observed in images of both formats: JPEG and FITS. In order to verify existence of this undesired feature across the whole data set, we retrieved and inspected all JPEG and FITS files available in the period from March 1, 2007 to September 20, 2013 (see Table 1).

As a simple measure of brightness variations, we chose the value of mean intensity level of the image. Analysis of this value clearly reveals brightness instability among images. The mean value of intensity level was calculated for each image, and results are represented in Figure 1. These plots led us to the conclusion that beacon data sets (both formats) consist of images with two types of intensity levels: low and normal. Some outliers that represent corrupted images are also noticeable on the plots.

Data set with JPEG images has much higher variability with bigger difference between STEREO A/B data. We also noticed that severe intensity level variability in the data set with JPEG images started on April 13, 2010, while FITS images have been staying mostly the same over this period of time. Based on these observations, we decided to focus on FITS images as the primary input to our segmentation algorithm.

Further analysis of the data set with FITS images revealed that the dynamic range of intensity levels was different between STEREO A/B images. The plots in Figure 2 show the difference between STEREO A/B FITS data sets. Each image is represented with 2 values: red square – minimum intensity level, and blue – maximum intensity level.

Although minimum intensity levels are almost the same for both COR2 instruments, the maximum intensity levels are different. The dynamic range for STEREO B is much higher that it creates additional challenges for development of generic automated image processing algorithm.

Table 1 Amount of files for the period from March 1, 2007 to September 20, 2013

Image Format	STEREO A Amount	STEREO B Amount	Totals	
FITS	98,299	104,129	202,428	382,947
JPEG	95,022	85,497	180,519	

3 Method

In this section we provide an overview of the proposed method, necessary pre-processing step, and data used in our validation..

3.1 Data

The data used for CME segmentation are STEREO A/B COR2 sixteen-bit grayscale FITS images. These data are collected by several NOAA and international ground-tracking stations. Data sets can be accessed using public and scientific STEREO Web pages at STEREO Science Center [4].

3.2 Pre-processing

The input to our method consists of series of twenty to thirty images. In most cases this amount of images ensures that there are images without CME in the series, since, on average, CME can be observed at one location on about ten consequent images. When generating mean image from the stack, non-CME pixels deemphasize regions that might represent CME. Mean image also, in some extent, collects noise patterns of each image from the stack and becomes efficient for background removal and CME segmentation.

In order to identify CME segments suitable for leading edge detection, an algorithm has to ensure that the dataset is homogeneous. Based on our analysis of the images, the data set suffers from intensity level instability. Not all series of images are affected by this issue and the role of the pre-processing step is to determine the presence of brightness variations. If it is present, then all images in the series are classified into two homogeneous classes. If not, then series of images is considered as one uniform class.

In order to perform classification, the mean intensity value is computed from the stack of all images in the data set.

The image is classified as “low class,” if its mean intensity value is lower than the mean intensity value of the data set; otherwise, it belongs to the “normal class”. Subsequently, two mean images are computed using stack of images of each class.

3.3 Segmentation

Berghmans et al stated that it turned out not to be feasible to identify in each separate image the location of individual CMEs by segmentation techniques. [13]. In agreement with this statement and our own experience we do not attempt to identify location of CMEs in separate images, but instead follow approach to segment CME masses by processing the series of images.

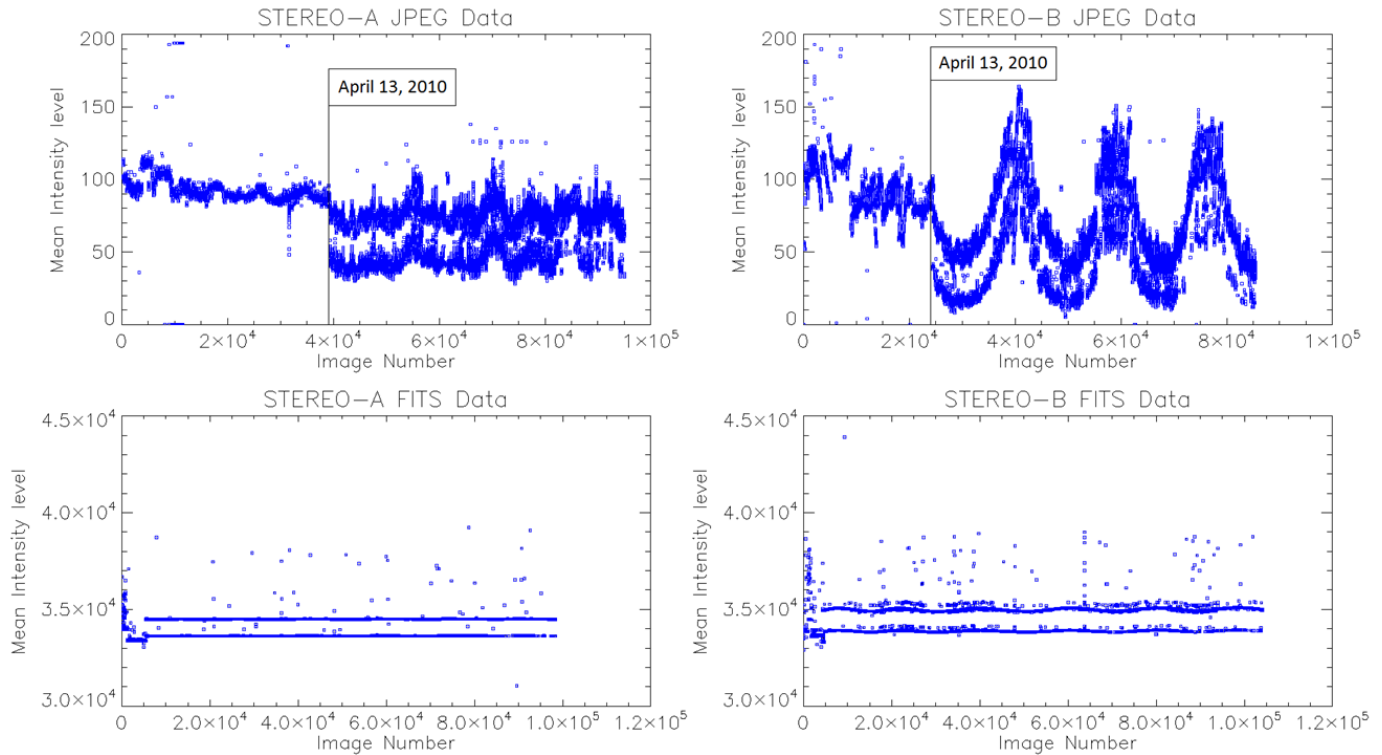


Figure 1 Mean intensity values for JPEG and FITS beacon images for the period from March 1, 2007 to September 20, 2013.

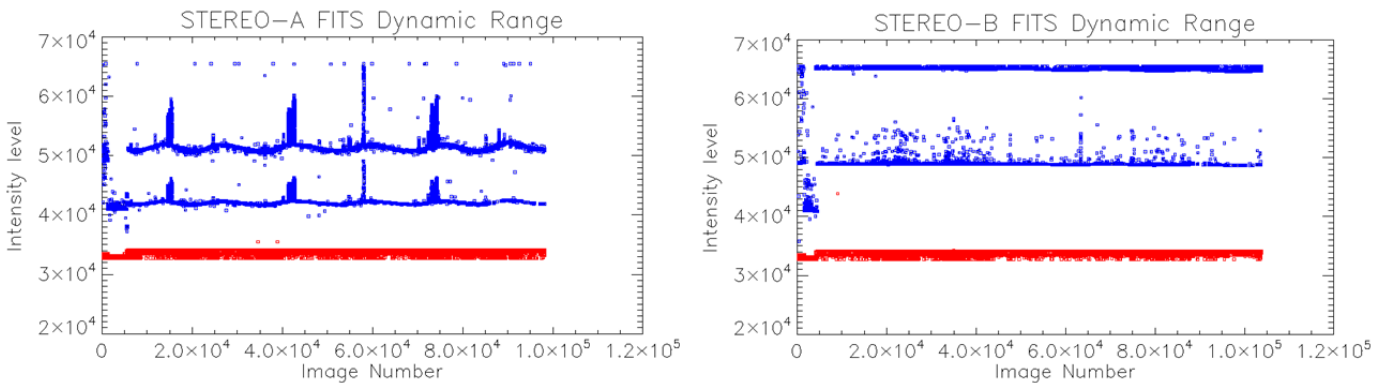


Figure 2 Dynamic ranges for STEREO A (left) and STEREO B (right) FITS images. Red - minimum intensity level. Blue – maximum intensity level.

During this step the corresponding mean image is subtracted from each image in the matching class leading to

the removal of background, leaving only the signal that is unique to each separate image. Images with segmented results

for both classes are merged into one group. At this point, some noise is still observed in the images in form of tiny connected regions or single isolated pixels with low brightness. Segments that refer to CMEs, in turn, have high brightness and bigger connected regions.

Morphological erosion and histogram equalization is applied to the images in order to remove remaining noise, increase contrast and equalize brightness. During the procedure of morphological erosion, segments that are smaller than the structuring element (size 3x3) get deleted. The histogram equalization procedure ensures equal distribution of intensities in every image.

3.4 Validation

To validate our approach, we used a data set of fifteen CME events, with series of FITS COR2 images from STEREO A/B spacecraft. In total the data set consists of thirty series, with twenty to thirty images each. The series of images were compiled manually to ensure that both STEREO A/B recorded these events in full. The validation was performed by domain experts who visually examined the segmented CMEs, and compared with the CMEs in original images.

4 Results

When inspected visually, was determined that all CME masses were segmented properly. Comparison of CMEs in original images with segmentation results revealed that the method segments CME masses successfully up to the distance of 12 solar radii using the selected 15 CME events.

Figures 3 and 4 show typical examples of segmentation result for CME event observed on 2013-04-05 by both spacecrafts. Red-colored images on the left are JPEG beacon images. These JPEG images are usually used for manual CME detection and provided here for visual comparison. It is clear that red-colored STEREO A images differ from STEREO B in brightness and contrast. In addition, STEREO B series includes class of low intensity images, which have the following timestamps: 07-39-24, 08-39-24, 09-39-24, 10-39-24, and 11-39-24. All defects that are observed in JPEG images exist in FITS as well.

The grayscale images on the right of Figures 3 and 4 represent results of segmentation. These images show the propagation of the CME up to 12 solar radii. As seen from these results, background is removed, the images are free from noise, and the CME is segmented successfully from all images in the series, including low intensity.

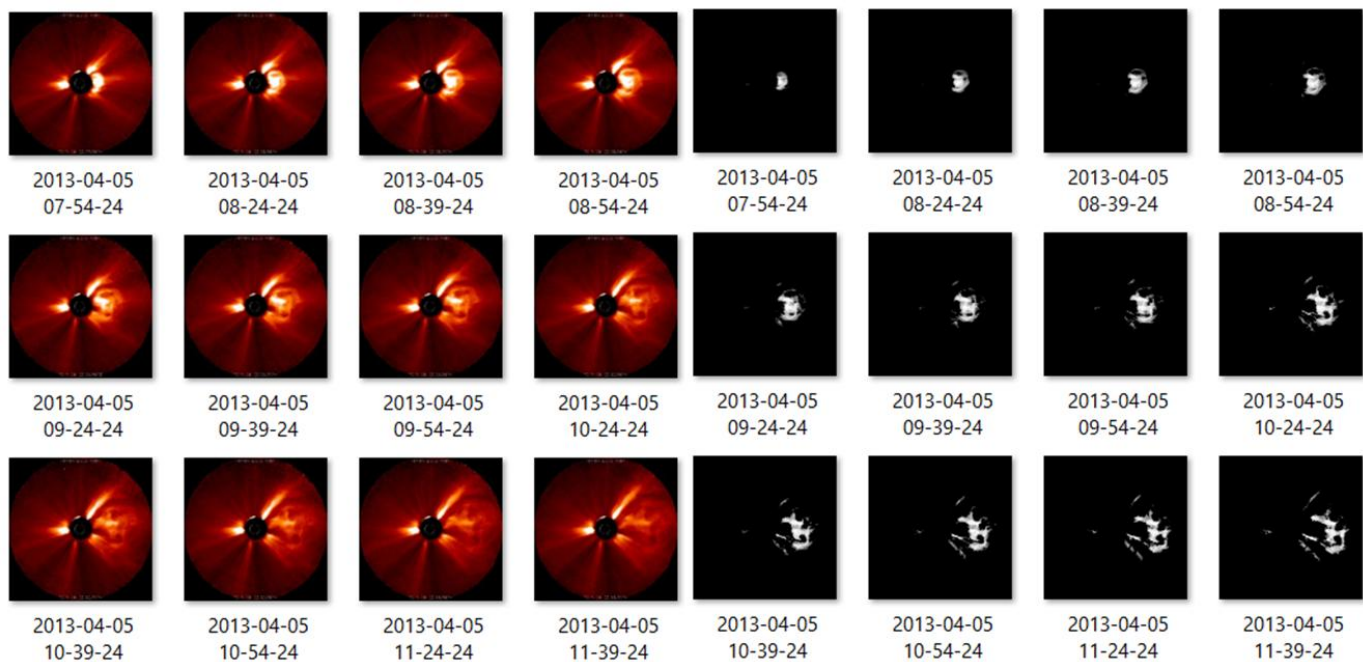


Figure 3 Segmented CME using the proposed method. The corresponding JPEG images from STEREO A (the first four columns) are given here for visual comparison.

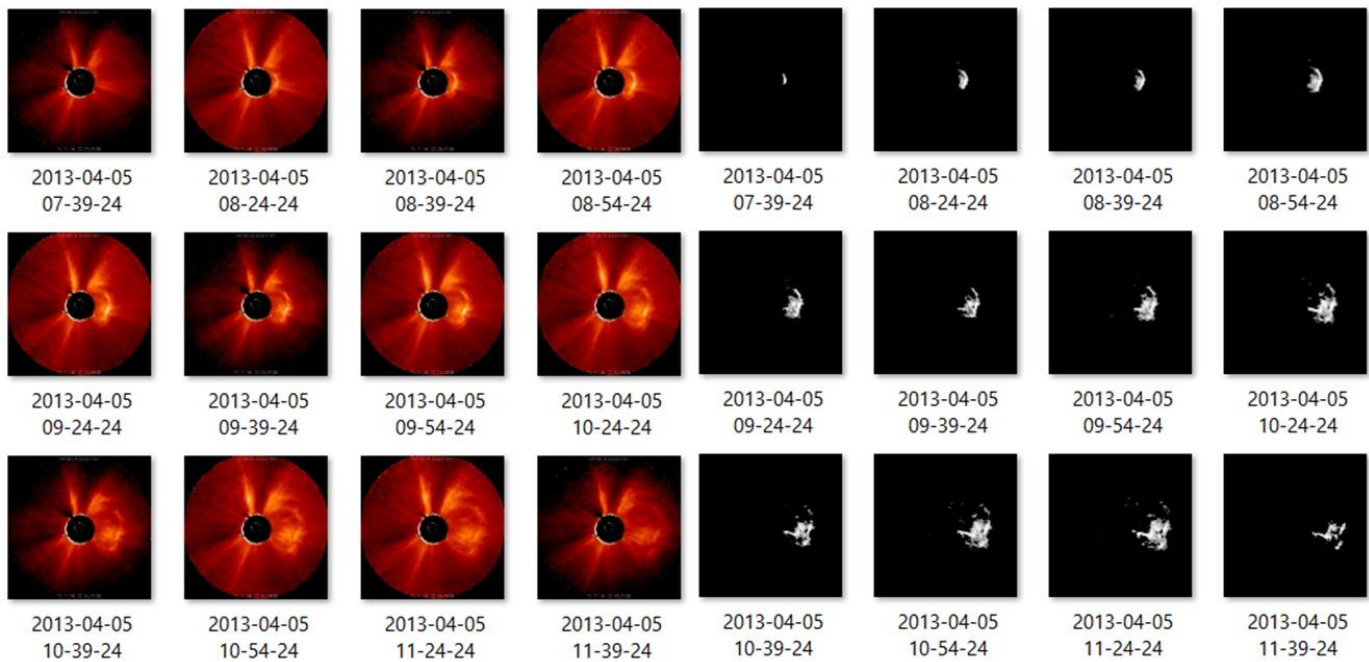


Figure 4 Segmented CME using the proposed method. The corresponding JPEG images from STEREO B (the first four columns) are given here for visual comparison.

5 Discussion and Conclusion

It was observed that FITS beacon COR2 data consist of images that vary in brightness. The proposed method pre-processed the data by dividing the input data into two classes. This pre-processing step is a key factor to overcome the signal instability issue. Mean image of each class is generated from homogeneous data since images in each class have similar brightness.

Segmentation is done by subtraction of relevant mean image from each image in the same class. Subtraction leads to background removal and significant noise suppression, leaving only signal that is unique to the image. Morphological erosion deletes remaining noise patterns and histogram equalization increases the contrast of the images, thus making CME masses more prominent.

The visual inspection has shown that the method segments images successfully and can be used as a data provider for further automatic classification, CME leading edge detection, and propagation parameters derivation.

In conclusion, this paper presented results of data set analysis that revealed some challenges in automatic CME segmentation and proposed a pre-processing step that helped to form the basis for development of the segmentation method. It also presented a novel method for segmentation of CMEs using STEREO beacon white-light coronagraph images. The method is generic and allows for processing

COR2 images in FITS format transmitted from both STEREO A/B spacecraft.

The experiments have demonstrated efficiency of the segmentation technique up to the distance of 12 solar radii that enable moving toward development of automatic methods for CME leading edge detection and extraction of propagation parameters.

6 Acknowledgement

The STEREO/SECCHI data used here are produced by an international consortium of the Naval Research Laboratory (USA), Lockheed Martin Solar and Astrophysics Lab (USA), NASA Goddard Space Flight Center (USA) Rutherford Appleton Laboratory (UK), University of Birmingham (UK), Max-Planck-Institut für Sonnensystemforschung (Germany), Centre Spatiale de Liège (Belgium), Institut d'Optique Théorique et Appliqué (France), Institut d'Astrophysique Spatiale (France).

7 References

- [1] Liu, Y.; Davies, J.A.; Luhmann, J.G.; et al. "Geometric triangulation of imaging observations to track coronal mass ejections continuously out to 1 AU"; *Astrophys. J. Lett.*, 710, 1, L82-L87, 2010.
- [2] Webb, D. F.; Howard, T. A. "Coronal Mass Ejections: Observations"; *Living Rev. Solar Phys.*, 9, 3, 2012.

- [3] Jacobs, M.; Chang, L.; Pulkkinen, A. "Automatic Segmentation and Classification of Multiple Coronal Mass Ejections from Coronagraph Images"; IPCV'13, 2013.
- [4] Kaiser, M. L.; Kucera, T. A.; Davila, J. M.; et al. "The STEREO Mission: An Introduction"; *Space Sci. Rev.*, 136, 1-4, 5-16, 2008.
- [5] Thernisien, A.; Vourlidas, A.; Howard, R. A. "CME reconstruction: Pre-STEREO and STEREO era"; *JASTP*, 73, 1156-1165, 2011.
- [6] NASA STEREO Learning Center. Beacon data. http://stereo.gsfc.nasa.gov/artifacts/artifacts_beacon.shtml
- [7] Robbrecht, E.; Berghmans, D. "Automated recognition of coronal mass ejections (CMEs) in near-real-time data"; *Astron. Astrophys.*, 425, 1097-1106, 2004.
- [8] Manual catalogs of CMEs. a) CDAW. http://cdaw.gsfc.nasa.gov/CME_list/index.html; b) DONKI. <http://swc.gsfc.nasa.gov/main/donki>; c) NRL. <http://lasco-www.nrl.navy.mil/index.php?p=content/cmelist>; d) STEREO COR1 Observers Log. <http://cor1.gsfc.nasa.gov/catalog/>.
- [9] Robbrecht, E.; Berghmans, D.; Van Der Linden, R. A. M. "Automated LASCO CME catalog for solar cycle 23: Are CMEs scale invariant?"; *Astrophys. J.*, 691, 1222-1234, 2009.
- [10] Boursier, Y.; Lamy, P.; Llebaria, A; et al. "The Artemis catalog of LASCO Coronal Mass Ejections. Automatic Recognition of Transient Events and Marseille Inventory from Synoptic maps"; *Solar Phys*, 257, 125-247, 2009.
- [11] Olmedo, O.; Zhang, J.; Wechsler, H; et al. "Automatic Detection and Tracking of Coronal Mass Ejections in Coronagraph Time Series"; *Solar Phys*, 248, 2, 485-499, 2008.
- [12] SEEDS catalog. On-line Web page. <http://spaceweather.gmu.edu/seeds/>.
- [13] Berghmans, D.; Foing, B. H.; Fleck, B. "Automated Detection of CMEs in LASCO Data"; *ESA SP*, 508, 437-440, 2002.

KPCA-based Node Selection for Fast KMSE

Jinghua Wang, Jane You, Qin Li

Department of Computing, The Hong Kong Polytechnic University,
Kowloon, Hong Kong

ABSTRACT

In this paper we first show that the kernel minimum squared error model is not computationally efficient in feature extraction. To speed up the feature extraction, we linearly express the feature extractor using nodes, i.e. a portion of the training samples in the kernel space. For node selection from the training set, we define two criteria based on Kernel principal component analysis. The nodes are representative and not similar to each other. The experimental results show the feasibility of the proposed method.

Keywords: Minimum Squared Error, Kernel Minimum Squared Error, Feature Extraction, Pattern Recognition

1. INTRODUCTION

Minimum Squared Error (MSE) [1,2] is a popular linear model for feature extraction and classification. The Kernel Minimum Squared Error (KMSE) method [3] is well known for its ability in nonlinear feature extraction. It is proved that KMSE is closely related to Kernel regression, Kernel Fisher discriminant analysis (KFDA) [4], and Support Vector Machine [3]. With the help of Kernel trick, we are able to implicitly map the samples into a high dimension space and transform the nonlinear problem into a linear problem [6,7,8,9]. Specifically, Kernel principal component analysis (KPCA) [15,16] is the nonlinear version of principal component analysis (PCA) [17].

Most kernel methods are proposed based on the Mercer's theory, i.e. the feature extractor in the kernel space is linearly expressible using the training samples. In the calculation of a single feature, we need to calculate as many kernel functions as the training samples. With this characteristic, the kernel methods are not applicable when the training set is very large. Some methods are proposed to accelerate the feature extraction procedure of the KMSE [5,12,13,14] and KFDA [10,18].

The idea of node is widely used to speed up the nonlinear feature extraction [5,6,10,12,13]. These methods linearly express the feature extractor using a portion of the training samples. The node-based methods work well mainly because some training samples have much larger coefficients than the other samples, i.e. some training samples are more important than the other samples. If we know the important samples, we can linearly express the feature extractor using them.

In this paper, we propose a new method for accelerating the feature extraction procedure of KMSE based on the idea of

KPCA. KPCA aims to extract the features by projecting samples on the most representative feature extractors, i.e. the eigenvectors of the covariance matrix. The representative ability of the feature extractors are assessed using the corresponding eigenvalues. In this paper, we select the most representative samples based on the idea of KPCA, and use these representative samples to linearly express the feature extractor. For simplicity, we call these representative samples as nodes. We define two criteria in this paper for node selection: 1) representative ability and 2) similarity. The first criterion assures that all of the nodes are as representative as possible. The second criterion assures that the node set is as small as possible.

Without consideration of the one-time node selection procedure, the proposed method is more computationally efficient than the naïve KMSE: 1) While the naïve KMSE needs to calculate the inverse matrix of size n by n , the proposed method calculate an inverse matrix of size h by h , where n and h are respectively the number of training samples and that of nodes. 2) In order to extract a feature, KMSE and the proposed method respectively need to calculate n and h kernel functions.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 presents the proposed method. Section 4 conducts experiments. Section 5 concludes this paper.

2. RELATED WORK

2.1 Minimum Squared Error (MSE)

For simplicity, in this paper we consider only binary classification problems. The following is the model of MSE

$$\begin{cases} X^+ w = c_1 \\ X^- w = c_2 \end{cases} \quad (1)$$

where $X^+ = [x_1^+ \ x_2^+ \ \cdots \ x_{n^+}^+]^T$ consists of the positive

samples and $X^- = [x_1^- \ x_2^- \ \cdots \ x_{n^-}^-]^T$ consists of the negative

samples. The elements of the vectors c_1 and c_2 are respectively composed of the class labels of the two classes. Normally, the elements of c_1 are all -1 while c_2 are all 1.

Denote all the training samples as $X = \begin{bmatrix} X^+ \\ X^- \end{bmatrix}$ and the labels as $C = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$, we can rewrite (1) as

$$XW = C \quad (2)$$

The equation (1) and (2) can be solved using the least squared-error technique. In a classification problem, we first extract a feature for the sample x using $x^T w$; then compare $x^T w$ with c_1 and c_2 . If the distance between $x^T w$ and c_1 is smaller than the distance between $x^T w$ and c_2 , we classify the sample x into the positive class; otherwise, we classify it into the negative class.

2.2 Kernel Minimum Squared Error (KMSE)

In KMSE, the original sample space is supposed to be mapped into a high dimensional space by a nonlinear function ϕ . The training samples in the kernel space are $\phi(x_1), \phi(x_2), \dots, \phi(x_n)$, where $n = n^+ + n^-$ is the total number of training samples. Corresponding the nonlinear mapping, there is a kernel function $k(x, y)$ to calculate the inner product of any two mapping samples $\phi(x_i)$ and $\phi(x_j)$ as follows:

$$k(x_i, x_j) = \phi^T(x_j) \phi(x_i) \quad (3)$$

The following is the mathematical model of KMSE [5]:

$$\Phi W = C \quad (4)$$

where

$$W = \begin{bmatrix} w_0 \\ w \end{bmatrix}, C = \begin{bmatrix} 1 \\ \vdots \\ -1 \end{bmatrix}, \Phi = \begin{bmatrix} 1 & \phi(x_1)^T \\ 1 & \phi(x_2)^T \\ \vdots & \vdots \\ 1 & \phi(x_n)^T \end{bmatrix} \quad (5)$$

The feature extractor W consists of a threshold w_0 and the vector w in the kernel space. Note, the dimensionality of w equals that of the training samples $\phi(x_1), \phi(x_2), \dots, \phi(x_n)$. The element $C_i (1 \leq i \leq n)$ is the class label of the i th sample $\phi(x_i) (1 \leq i \leq n)$. KMSE can be considered as a machine, where $\phi(x_i)$ is the input and its label C_i is the output. Based on the Mercer's theory, we can express the feature extractor as follows [3]:

$$W = \begin{bmatrix} w_0 \\ \sum_{i=1}^n \alpha_i \phi(x_i) \end{bmatrix} \quad (6)$$

Substituting equation (6) into equation (4), we obtain:

$$KA = C \quad (7)$$

where

$$A = \begin{bmatrix} w_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, K = \begin{bmatrix} 1 & k(x_1, x_1) & \cdots & k(x_1, x_n) \\ 1 & k(x_2, x_1) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \quad (8)$$

Equation (9) is the normal equation of (7) and they have the same solution:

$$(K^T K) A = K^T C \quad (9)$$

If the matrix K is a full column rank matrix, the solution is:

$$A = (K^T K)^{-1} K^T C \quad (10)$$

If the matrix $K^T K$ is singular, equation (11) calculates a numerical stable solution of equation (7):

$$A = (K^T K + \mu I)^{-1} K^T C \quad (11)$$

where μ is a positive constant and I is the identity matrix [16].

For a test sample x , KMSE extracts a feature by projecting its mapping sample $\phi(x)$ onto the feature extractor W , as follows

$$l_p(x) = w_0 + \sum_{i=1}^n \alpha_i k(x, x_i) \quad (12)$$

If $l_p(x) > 0$, we label x with 1; or else, we label it with -1.

2.3 Kernel Principal Component Analysis (KPCA)

Here, we assume the mapping samples are centered. Then, we can calculate the covariance matrix as follows

$$S_i^\phi = \sum_{i=1}^n \phi(x_i) \phi^T(x_i) = X X^T \quad (13)$$

where $X = [\phi(x_1) \ \phi(x_2) \ \cdots \ \phi(x_n)]$ is a matrix consists of all the mapping samples. KPCA aims to extract the most representative features and the feature extractors are the eigenvectors of the following eigenequation

$$S_i^\phi v = \lambda v \quad (14)$$

The importance of a feature extractor is measured by the corresponding eigenvalue. To achieve the goal of dimension reduction, we usually only keep the most important feature extractors. These feature extractors are linearly expressed by all of the training samples [11], i.e.

$$v = \sum_{i=1}^n \alpha_i \phi(x_i) = X \alpha \quad (15)$$

where $\alpha = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_n]^T$ is a vector in n dimensional space. With equation (14) and (15), we obtain the flowing matrix

$$P P \alpha = \lambda P \alpha \Rightarrow P \alpha = \lambda \alpha \quad (16)$$

where the matrix $P = X^T X$ is defined to be

$$P_{ij} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j) (1 \leq i, j \leq n).$$

3. PROPOSED METHOD

As can be seen from equation (12), we need to calculate the kernel function between x and every training sample. Thus, the feature extraction efficiency of KMSE is inversely proportional to the size of the training set. This means KMSE is not applicable in large scale problems. It is necessary to improve KMSE for fast feature extraction.

In the proposed method, we assume the feature extractor of KMSE can be linearly approximated by nodes, i.e. a portion of important training samples. We have such an assumption based on the observation that some training samples have much larger coefficient than the rest. To approximate the feature extractor as well as possible, the nodes should be as representative as possible. If we can use these nodes to linearly approximate the training samples which are not nodes, we can use the node set to replace the training set in the linear expression of the feature extractor. Before presenting the criteria for node selection in section 3.2, we first show the formulation of the proposed method in section 3.1.

3.1 The proposed method

Suppose the feature extractor can be approximated as follows

$$W \approx W^* = \begin{bmatrix} w_0 \\ \sum_{i=1}^h \beta_i \phi(x_i^*) \end{bmatrix} \quad (17)$$

where $x_1^*, x_2^*, \dots, x_h^*$ are the nodes and $\beta = [\beta_1 \ \beta_2 \ \dots \ \beta_h]^T$ is a h by 1 vector. In the following, we replace the sign “ \approx ” in (17) with “ $=$ ”. By substituting (17) into (4), we obtain

$$K^* A^* = C \quad (18)$$

where

$$A^* = \begin{bmatrix} w_0 \\ \beta_1 \\ \vdots \\ \beta_h \end{bmatrix}, K^* = \begin{bmatrix} 1 & k(x_1, x_1^*) & \dots & k(x_1, x_h^*) \\ 1 & k(x_2, x_1^*) & \dots & k(x_2, x_h^*) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k(x_n, x_1^*) & \dots & k(x_n, x_h^*) \end{bmatrix} \quad (19)$$

A similar equation to (11) is used to calculate the coefficient vector A^* , as follows

$$A = (K^{*T} K^* + \mu I)^{-1} K^{*T} C \quad (20)$$

For a test sample x , the proposed method extracts a feature by projecting its mapping sample $\phi(x)$ onto the feature extractor W^* , as follows

$$l_p(x) = w_0 + \sum_{i=1}^h \beta_i k(x, x_i^*) \quad (21)$$

If $l_p(x) > 0$, we label x with 1; or else, we label it with -1 .

In (21), we only need to calculate a kernel function for each of nodes. In our experiments, the nodes are only a small portion of the training samples. Thus, the feature extraction procedure in (21) is much more computationally efficient than that in (12). In addition, the equation (20) calculates an inverse matrix of a matrix of size h by h . Differently, equation (11) has to calculate an inverse matrix of size n by n . So, the calculation of the coefficient vector in the proposed method is more computationally efficient.

3.2 Node selection

The performance of the proposed method depends on the node set. Generally speaking, the more representative the node set is, the better we can use the nodes to approximate the feature extractor. Here, we employ the idea of KPCA. In KPCA, the importance of an eigenvector is assessed using the corresponding eigenvalue. Similarly, we define a criterion (pseudo-eigenvalue) l to assess the training sample $\phi(x)$ as follows

$$l = \frac{\phi^T(x) S_i^o \phi(x)}{\phi^T(x) \phi(x)} \quad (22)$$

The larger the value l , the more representative the training sample is. While a matrix S_i^o only has n eigenvalues, it has a pseudo-eigenvalue for each of the training samples.

In order to reduce the size of the node set, we bound the similarity between the selected nodes using the following criteria

$$\cos(x_i, x_j) = \frac{\phi^T(x_i) \phi(x_j)}{\sqrt{\phi^T(x_i) \phi(x_i)} \sqrt{\phi^T(x_j) \phi(x_j)}} \quad (23)$$

If a sample is selected as a node, all the other samples which are similar to it are deleted from the candidate set.

The following is the node selection algorithm

- Step 1 initialize the node set to be null and the candidate set to be the training set;
- Step 2 put the most representative sample into

the node set;

- Step 3 delete each sample from the candidate set if its similarity with the node is higher than a threshold;
- Step 4 if the candidate set is not null, go to step 2.

With the selected node, we can formulate the proposed method according to (18) and extract features using (21).

4. EXPERIMENTS

Our experiments are conducted using several benchmark datasets downloaded from <http://archive.ics.uci.edu/ml/>. For each dataset, we randomly separate it into training set and testing set and repeat ten times. We use the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$. The parameter σ^2 is automatically set to be the sum of the variances of all the data components in the training sample subset. The parameter μ equals to 0.1 for of the datasets. We conduct experiments using the ten divisions of the sample set. The average of ten classification accuracy is regarded as final accuracy in our experiments.

We compare the proposed method with the standard KMSE and its three improvements, RKMSE[12], DKMSE[13], MPLOC-KMSE[14], and EKMSE[5]. All of the four methods speed up the feature extraction procedure by the idea of node, i.e. expressing the feature extractor using a portion of the training samples.

Table 1 the number of nodes in different methods

dataset	Titanic	flare	Heart	German
(number of samples)	150	666	170	700
RKMSE[12]	26	49	64	238
DKMSE[13]	17	51	70	180
MPLOC-KMSE [14]	7	14	90	206
EKMSE[5]	8	19	47	186
Proposed method	5	9	39	169

Table 2 error rates of different methods (%)

dataset	Titanic	flare	Heart	German
KMSE	22.6	34.3	12.0	21.7
RKMSE[12]	26.4	34.7	13.8	28.4
DKMSE[13]	25.2	32.5	14.7	27.5
MPLOC-KMSE [14]	25.6	33.2	13.4	26.4
EKMSE[5]	24.1	31.2	13.0	21.3
Proposed method	23.4	30.9	12.4	19.7

5. CONCLUSION

MSE and KMSE are extensively used in application. However, the feature extraction procedure of the KMSE is inversely proportional to the size of the training set. In this paper, we reformulate the KMSE by linearly express the feature extractor as a linear combination of a portion of the training set. This speeds up the calculation of the coefficient vector as well as the feature extraction procedure. The experimental results show the feasibility of the proposed method.

6. REFERENCES

- [1] K.R. Muller, S. Mika, G. Ratsch.,K. Tsuda, & B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Trans. On Neural Network, 12(1), 2001, 181-201.
- [2] J. Yang, Z. Jin, J.Y. Yang, D. Zhang, & A.F. Frangi, Essence of kernel Fisher discriminant: KPCA plus LDA, Pattern Recognition, 37(10), 2004, 2097-2100.
- [3] J. Xu, X. Zhang, & Y. Li, Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR. Proc. of the International Joint Conference on Neural Networks (IJCNN-2001), Washington, D.C, 2001, 1486-1491.
- [4] S. Mika, G. Rätsch, and J. Weston, et al., Fisher discriminant analysis with kernels, In: Y.H. Hu, J. Larsen, E. Wilson, & S. Douglas, Neural Networks for Signal Processing IX (IEEE, 1999, 41-48).
- [5] J. Wang, P. Wang, Q. Li, J. You, Improvement of the kernel minimum squared error model for fast feature extraction, 23 (1), 2013, 53-59.
- [6] Y. Xu, J.Y. Yang, J. Lu, & D.J. Yu, An efficient renovation on kernel Fisher discriminant analysis and face recognition experiments, Pattern Recognition, 37(10), 2004, 2091-2094.
- [7] Y.C. Eldar, & T.G. Dvorkind, A minimum squared-Error framework for generalized sampling, IEEE Transactions on Signal Processing, 54(6), 2006, 2155-2167.
- [8] R.O. Duda, P.E. Hart, & D.G. Stork, Pattern Classification (second edition)(Beijing: China Machine Press, 2004).
- [9] F.X. Song, J.Y. Yang, & S.H. Liu, Pattern recognition based on the minimum norm minimum-squared error classifier, The Eighth International Conference on Control, Automation, Robotics and Vision, Kunming, China, 2004, 1114-1117.
- [10] J. Wang, Q. Li, J. You, Q. Zhao, Fast Kernel Fisher Discriminant Analysis via Approximating the Kernel Principal Component Analysis, Neurocomputing, 74(17), 2011, 3313-3322.
- [11] B. Scholkopf, A. Smola, K.R. Muller, Kernel principal component analysis, Artificial Neural Networks-ICANN'97, Berlin, 1997, pp. 583-588.
- [12] Q. Zhu, Reformative nonlinear feature extraction using kernel MSE, Neurocomputing 73(16-18) (2011) 3334-3337.
- [13] Y. Xu, J.-Y. Yang, J.-F. Lu, An efficient kernel-based nonlinear regression method for two-class classification, in Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, 2005, vol. 7, pp. 4442-4445.
- [14] Y.-P. Zhao, J.-G. Sun, Z.-H. Du, Z.-A. Zhang, H.-B. Zhang. Pruning least objective contribution in KMSE. Neurocomputing 74(17)(2011) 3009-3018.

- [15] K.I. Kim, K. Jung, H.J. Kim, Face recognition using kernel principal component analysis, *IEEE Signal Processing Letters* 9 (2) (2002) 40-42.
- [16] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* 10 (5) (1998) 1299-1319.
- [17] K. Fukunaga, *Introduction to statistical pattern recognition*, second edition, Academic Press, Inc., New York, 1990.
- [18] A.J. Smola, B. Scholkopf, Sparse greedy matrix approximation for machine learning, in: *Proceedings of the 17th International Conf. on Machine Learning*, San Francisco, 2000, pp. 911-918.

Liver Extraction from CT Images Based on Liver Structure Models

Masanori Hariyama¹, Riichi Tanizawa¹, Mitsugi Shimoda², Keiichi Kubota²

¹Graduate School of Information Sciences, Tohoku University, Japan

² Second Department of Surgery, Dokkyo Medical University

Abstract—The extraction of a liver from CT images is essential for oncologic surgery planning. This article presents an accurate and automatic approach to extract a liver from CT images. Our algorithm exploits three types of liver structure models: intensity model, shape model, and blood vessel model. First, the region including the liver is roughly extracted based on intensity histogram analysis. Second, the extracted regions are segmented using a shape feature called “local thickness” based on the observation that the liver is thicker than other organs. Finally, the segmented regions including blood vessels in the liver are merged into a single liver region. Experimental results show that the average error of the volume extraction is 61.25 cc, and this result is much superior to the conventional one.

Keywords: Medical imaging, 3D simulation analysis, anatomic hepatectomy, local thickness

1. Introduction

For liver cancer surgery, 3D simulation before surgery operations, recently, is getting one of the crucial tasks since a liver has complex structure. Liver segmentation is considered as a challenging task since the variations of the liver shape is large and since there exist some other organs with the CT values similar to the liver around the liver. There have been several researches on liver extraction[2]-[5]. The researches [2] and [3] use, respectively, statistical shape models and probabilistic atlases, and both methods suffers from large variations of liver shapes. The active contour approach [4] are dependent on image gradient, and leads to over-extraction into organs with CT values similar to the liver. Moreover, its quality strongly relies on the location and shape of the initial contour. The intensity-based approach [5] usually exploits a simple intensity model, and miss the vessels and non-homogenous texture regions inside the liver.

This paper propose a new accurate intensity-based approach. In order to improve the quality, we use additional models: a shape and vessel models as well as an intensity one.

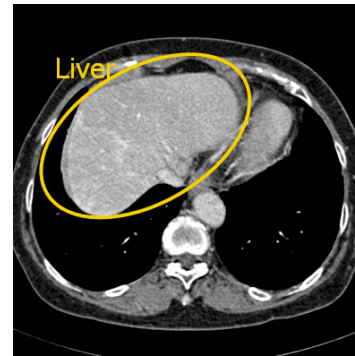


Fig. 1: CT image including the liver.

2. Liver structural model and extraction algorithm

2.1 Extraction of liver candidate regions based on an intensity model

Figure 1 shows a CT image of the liver. The liver is one of the biggest organs and the intensities of the liver points are evenly high. Figure 2 explains how the candidate regions of the liver are extracted based on the intensity model. First, the histogram of the intensity is computed as shown in the upper part of Fig. 2. The histogram has usually two mounts; the darker mount corresponds to the fat; the brighter one corresponds to the liver, spleen, born, etc. The brighter part is extracted automatically by using the Otsu’ thresholding method[1]. Next, for the resulting 3D image, the thickness feature is measured by computing “Local Thickness“ [6], where the local thickness of a point is defined as the diameter of the largest sphere that fits inside the object and contains the point, as shown in the lower part of Fig. 2. Since the liver region is large and thick, the region with largest local-thickness values is extracted as the core of the liver region. For this core region, the intensity histogram is computed to get accurate intensity thresholds for the liver region. Finally, the liver candidate regions are extracted by thresholding using the thresholds.

2.2 Segmentation based on a shape model

The most outstanding shape feature of the liver is that the liver surface is smoothly rounded and the liver is thick.

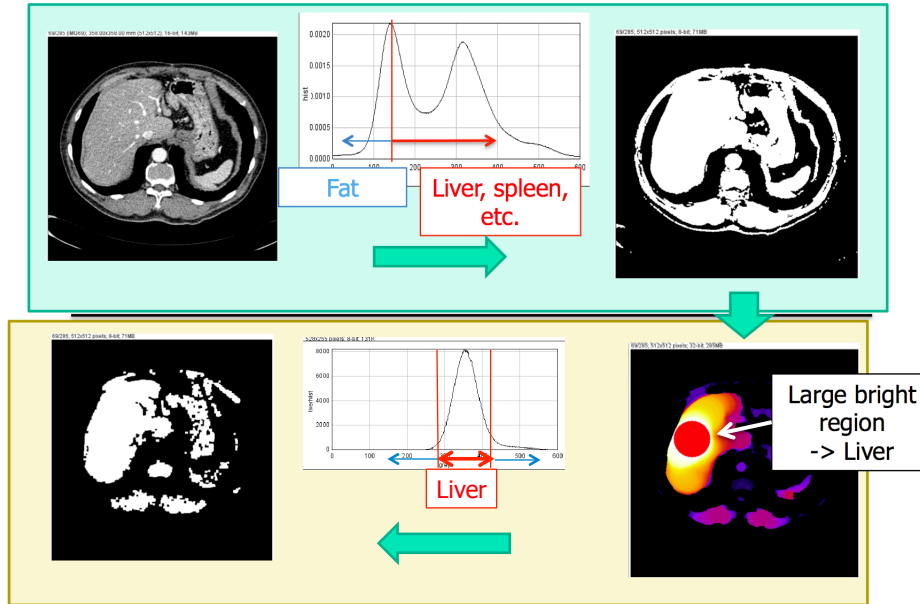


Fig. 2: Flow of extracting the liver candidate regions based on the intensity model.

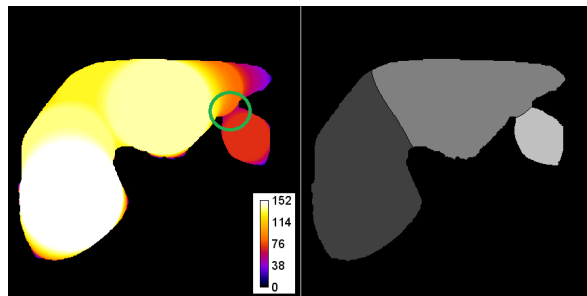


Fig. 3: Segmentation using a shape model.

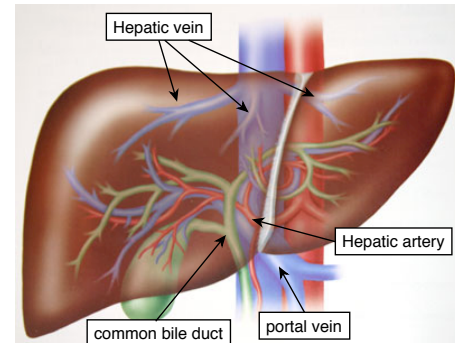


Fig. 4: Vascular system of the liver.

In order to measure the roundness and thickness at the same time, the local thickness is used. The 3D image (the result of Section 2.1) is segmented as follows. The local thickness values for all point in 3D image are computed as shown the left part in 3. The local thickness image is segmented using Watershed method[7], where the point with the maximum value is used as the seeds, and the point with minimum value is used as watershed points. The Watershed algorithm separates the different organs well since the local thickness tends to be minimum at the boundary points where different organs touches. Figure 3 shows an example of the watershed-based segmentation, where the local-thickness image is segmented into three regions (two regions of the liver and one region of the spleen). Although the Watershed algorithm might segment the local-thickness image into a lot of small regions, appropriate regions are picked up to be merged into a liver region by the process described in the next section.

2.3 Merging the segmented regions based on a vessel model

In a liver, there are three types of vessels: hepatic vein, portal vein, and hepatic artery as shown in Fig. 4. Based on this observation, the segmented regions in the previous process are merged into a single liver region by using the vessel information. If a segmented region touches the vessels, it is picked up for a liver region. The vessel data is obtained by using a vessel extraction program developed by Hariyama et al.; it extracts the vessels based on line filter[8] and refines the extraction result based on structural analysis. Figure 5 shows an example of this merging process.

3. Evaluation

Let us compare the proposed method with a conventional one where the liver region is automatically extracted and

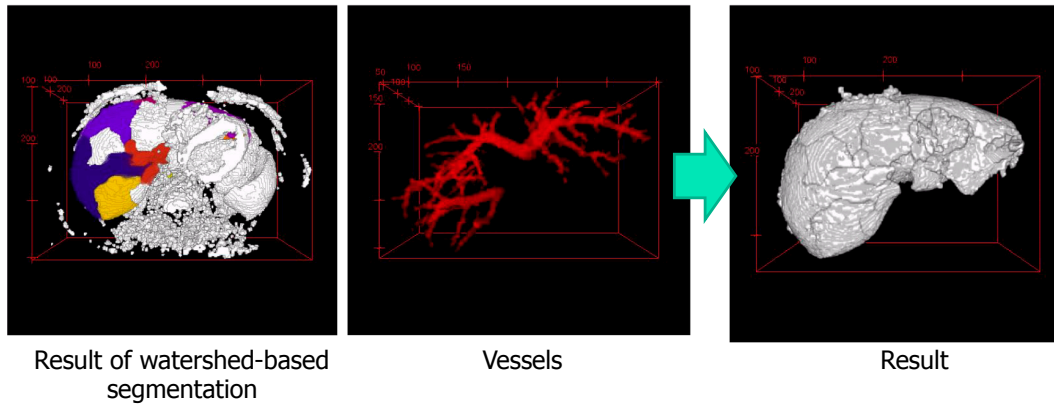


Fig. 5: Merging the segmented regions using a vessel model.

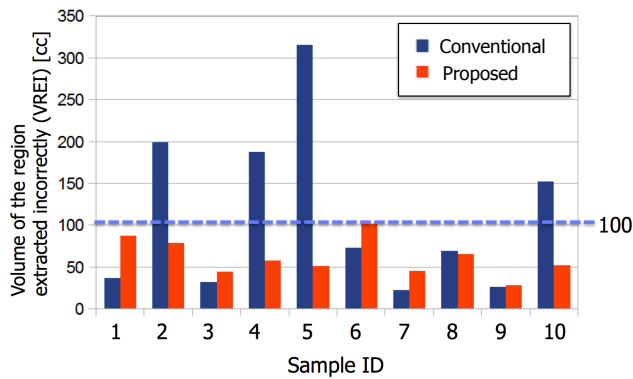
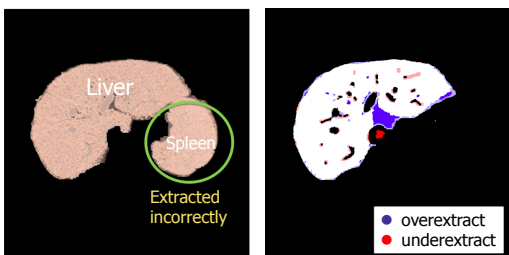
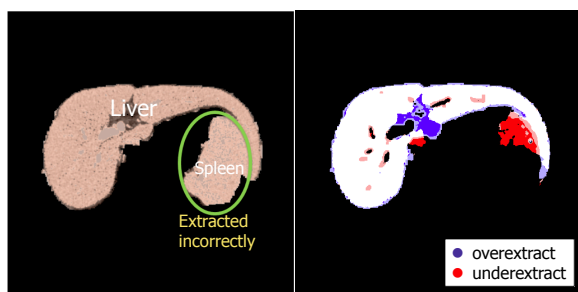


Fig. 6: Comparisons with the conventional method for 10 samples.



(a) Conventional (b) Proposed
Fig. 7: Improvement in sample 4.



(a) Conventional (b) Proposed
Fig. 8: Improvement in sample 10.

not modified manually. As a conventional software, Synapse VINCENT (Fujifilm Medical, Tokyo, Japan)[9] (ver. 2) is used which is one of the widely-used programs for 3D simulation.

The comparisons is done for 10 samples with grand truth which is made manually in terms of the volume of the region extracted incorrectly(called *VREI* in the following), which is the sum of the volumes of over-extracted and under-extracted volumes. Figure 6 summarizes the comparison results. The average VREIs of the proposed and conventional methods are 61 [cc] and 111 [cc], respectively. The standard deviations of the proposed and conventional methods are 20 and 89, respectively. The proposed method is more robust than the conventional one for change of samples. From some surgeons' experiences, the VREI less than 100 [cc] is desirable for preoperative planning. ;the proposed method can satisfy this requirements. Figures 7 and 8 compares the extracted liver regions for sample 4 and sample 10. In the conventional method, the spleen is extracted incorrectly as the liver since the spleen has a intensity feature similar to the liver. On the other hand, in the proposed method, the spleen is almost removed by exploiting the shape and vessel models.

4. Conclusion

The proposed method can improve the extraction accuracy by combining different types of features. The comparison results demonstrates that the proposed method is more robust for differences of patients. As future work, simultaneous recognition of other organs around the livers is on-going to improve the accuracy.

References

- [1] N. Otsu, "A threshold selection method from gray-level histograms", IEEE Trans. Sys., Man., Cyber. Vol. 9, Issue 1, pp.62-66 (1979).
- [2] Y. Song, A.J. Bulpitt, and K. Brodlie, " Liver segmentation using automatically defined patient specific B-Spline surface models, " MICCAI 2009 London, pp.43-50(2009).

- [3] G. Linguraru, J.K. Sandberg, Z. Li, F. Shah, and R.M. Summers, "Automated segmentation and quantification of liver and spleen from CT images using normalized probabilistic atlases and enhancement estimation," *Medical Physics*, vol.37, no.2, pp.771-783, 2010.
- [4] R.S. Alomari, S. Kompalli, and V. Chaudhary, "Segmentation of the liver from abdominal CT using Markov random field model and GVF snakes," *Proc. 2008 International Conference on Complex, Intelligent and Software Intensive Systems*, pp.293-298, 2008.
- [5] A.H. Foruzan, R.A. Zoroofi, M. Hori, and Y. Sato, "A knowledge-based technique for liver segmentation in CT data," *Computerized Medical Imaging and Graphics*, vol.33, no.8, pp.567-587(2009).
- [6] R. Dougherty and K. Kunzelmann, "Computing local thickness of 3D structures with ImageJ", *Proc. Microscopy & Microanalysis Meeting*, www.optinav.com/LocalThicknessEd.pdf (2007)
- [7] Vincent, L., Soille, P., "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *Pattern Analysis and Machine Intelligence*, IEEE Transactions on Volume 13, Issue 6, Jun 1991, pp.583 - 598.
- [8] Y. Sato, S. Nakajima, N. Shiraga, H. Atsumi, S. Yoshida, T. Koller, G. Gerig, and R. Kikinis, "Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images", *Med.Image Anal.*, vol2, no.2, pp.143-168(1998).
- [9] S. Ohshima, "Volume analyzer SYNAPSE VINCENT for liver analysis", *Journal of Hepato-Biliary-Pancreatic Sciences*, Vol. 21, Issue 4, pp. 235-238(2014)

Genetic Algorithm-Based Image Feature Matching with a Fundamental Matrix

Jaeyoung Kim¹, Myeongsu Kang¹, Heesung Jun¹, and Jong-Myon Kim^{1,*}

¹Department of Electrical, Electronics, and Computer Engineering, University of Ulsan, Ulsan, South Korea
kim7097@mail.ulsan.ac.kr, ilmareboy@ulsan.ac.kr, hsjun@ulsan.ac.kr, jmkim07@ulsan.ac.kr

Abstract - This paper proposes an efficient method to filter the incorrect matches of scale invariant feature transform (SIFT)-based feature matching which includes correct and incorrect matches using a fundamental matrix and a genetic algorithm. First, the best fundamental matrix is estimated by the genetic algorithm whose input data are unfiltered feature matches. To design chromosome, each gene represents the index of selected matches and it has a linkage value which is got from the linkage map. Selected matches are used to estimate fundamental matrix candidates. Each chromosome has a fitness value related to the number of found correct matches and matching score which is calculated from its fundamental matrix. To crossover, multiple order crossover operation is applied. For replacement, k -rank chromosomes are replaced from sorted global generation. Then, the best fundamental matrix is used for filtering. Finally, good matches are extracted by comparing RANSAC and LMED fundamental matrix estimation.

Keywords: Feature matching, fundamental matrix, genetic algorithm, SIFT

1 Introduction

In computer vision, a correspondence problem between two sets of points is very important and complex problem. It is related to many tasks such as object recognition, 3D reconstruction, image stitching and stereo matching. Scale invariant feature transform (SIFT) [1] is very robust in many kinds of transformation such as rotation, shifting, scaling, affine, perspective and illumination changes. However, if we only use the descriptor distances between two features to match features of query image to features of train image, it is not accurate since geometric information is not considered and many similar descriptors could exist. Many researches are presented for a decade. Random sample consensus (RANSAC) algorithm was used to estimate good homography matrix which is fit to given correct matches, and it is used to filter incorrect matches of original matches [2, 3].

However, this method is only robust for planar-transformation because the homography matrix can only transform point to point on each different plane. Spectral graph matching uses pairwise relationship between two matches [4], but it is not good in scale transformation because the distances between two features on the same image may be sensitive to scale [5]. Tensor-based hyper graph matching uses 3-tuple relationship which is triangular similarity [5], but it is not good in affine and perspective transformation because triangular similarity is very sensitive to these kinds of transformations.

In this paper, we propose an efficient method to find correct matches between two sets of features on the query image and train image. We use a fundamental matrix as geometric information because the fundamental matrix includes camera parameters as well as 3D geometric information. Thus, it is available to find not only planar correct matches but also 3D geometrical correct matches such as sphere and cube.

The rest of this paper is organized as follows. Section 2 briefly introduces background information of the SIFT algorithm, and Section 3 explains how to estimate fundamental matrix using a few selected matches. To estimate a good fundamental matrix with whole correct matches, we execute a genetic algorithm iteratively. Section 4 describes how to design chromosome representation and what kind of crossover operation is used. Section 5 shows experimental results for finding correct matches using best estimated fundamental matrix. In addition, we compare our method with the conventional RANSAC and least median square (LMEDS) which are implemented in OpenCV library [6]. Finally, Section 6 concludes this paper.

2 Scale invariant feature transform

The SIFT algorithm, which was proposed by David G. Lowe [1], extracts stable feature points and generates discriminant descriptors for each feature point based on the scale-space theorem and the gradient measurement. SIFT consists of four steps: scale-space peak detection, keypoint localization, orientation assignment, and local image descriptor generation.

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. NRF-2013F1A2A2A05004566).

* Corresponding author.

2.1 Scale-space peak detection

This step creates scale-space to find scale-invariant points called keypoints which can be frequently represented in different viewpoints. According to the Lindeberg's assumptions [7], Gaussian function is suitable as a kernel function of the Gaussian scale-space which is defined as:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (1)$$

where $G(x, y, \sigma)$ is a Gaussian function at point (x, y) of the image, σ represents the scale, operator $*$ is the convolution, and $I(x, y)$ is the intensity of the image at point (x, y) . To find stable keypoints, the difference of Gaussian (DoG) is generated by subtracting two closest scales in the scale-space. DoG is defined as follows:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma). \quad (2)$$

Lowe empirically defined both scaling value k as $\sqrt{2}$ and σ as 1.6 [1]. After the specific number of scales is calculated, octaves are created by down sampling the image and calculating DoG with doubled σ . In addition, the optimal number of scales per octave was experimentally set to three [1]. The local minimum and maximum points, which are candidate keypoints, are found from the DoG. To find them, the current point is compared with the eight neighbor points in the same scale. If it is a maximum or minimum point, the current point is compared with more 18 points which are located in up and down scale of the current scale. This process is iterated until all scales and octaves are scanned.

2.2 Keypoint localization

This step can be separated into three stages. The first stage is the interpolation of keypoint positions. In the previous step, the position of candidate keypoints is not accurate when it is applied to the original image since it is affected by the differential. To interpolate it, *Taylor* expansion is employed to express DoG such as

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x. \quad (3)$$

The interpolated keypoint position can be calculated by solving (3) for x ,

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x}. \quad (4)$$

The second stage is to discard low-cost keypoints. If the value of second-order *Taylor* expansion at \hat{x} is less than 0.03, the candidate keypoint is discarded. Otherwise, it is preserved. The last stage discards edge responses. *Hessian* matrix can be utilized to detect edge response,

$$H(x) = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}, \quad (5)$$

where D_{xx} is the second derivative of DoG for x .

SIFT calculates trace-determinant ratio, which is defined as $Tr(H)^2 / Det(H)$. If the trace-determinant ratio is less than $(r+1)^2 / r$, in which r is a predefined threshold eigenvalue ratio of *Hessian* matrix H , the candidate keypoint is considered responding an edge point and discarded. In this study, r is set to 10 [1].

2.3 Orientation assignment

For rationally invariance, the gradient information of the image is utilized. Gradient can be calculated by measuring the magnitude and orientation of the difference of four neighbor pixels. Both magnitude and orientation of the gradient information can be defined as follows:

$$m(x, y) = \sqrt{d_x^2 + d_y^2}, \quad (6)$$

$$\theta(x, y) = \tan^{-1}(d_y / d_x), \quad (7)$$

where $d_x = L(x+1, y) - L(x-1, y)$ and $d_y = L(x, y+1) - L(x, y-1)$.

2.4 Local image descriptor generation

This step is required to make each keypoint more discriminant in the change of 3D viewpoint or illumination. Each patch of local neighbor pixels which centers responding keypoint is sampled and calculates the gradient between closest neighbor points. The calculated gradient magnitude and orientation are used to create 4 x 4 array histograms, where each histogram consists of eight orientation bins. Each histogram is located at the position of the sampled pixels (i.e, a square including all samples). Thus, each keypoint has a 128-element vector, where each element is the bin of orientation histogram of the 4 x 4 array histograms.

3 Fundamental matrix estimation

Fundamental matrix is an algebraic representation of epipolar geometry. We can derive the fundamental matrix from the mapping between a point and its epipolar line, and then specify the properties of matrix [8]. In general, we need more than 7-point pairs to estimate fundamental matrix. The constraint is that one point must be in another plane at least because if all points in the same plane, the fundamental matrix only includes planar information. In our algorithm, an 8-point fundamental matrix estimation method is used. If we have more than 8 correct matches, we can calculate fundamental matrix by solving (8):

$$x_i^T F x = 0, \tag{8}$$

where F is a fundamental matrix corresponding to the eight selected matches $x_i \rightarrow x'_i$. After fundamental matrix estimation, Eq. (8) should be applied to all of original matches. If the left side of (8) is close to zero, the match approximately corresponds with the fundamental matrix, F . In our experiment, we choose the threshold value as 0.05 empirically. If the absolute value of the left side of (8) is larger than the threshold value, the matches are removed as outliers.

4 Genetic algorithm

In the proposed method, to estimate F matrix, more than eight correct matches should be selected among all matches and the fundamental matrix should be evaluated through the original matches. It is a complex NP-hard problem because the number of combination is ${}_m C_8$ and its time complexity is $O(m^8)$, where m is the number of original matches. The genetic algorithm is widely used as a very efficient method to solve the NP-hard problem. The proposed method uses the genetic algorithm which consists of selection, crossover and replacement, and its procedure is as follows:

- Generate N initial chromosome as first generation, where N is population size.
- Select two parents of current generation randomly.
- Crossover the parents to generate offspring.
- The offspring is put into the next generation and go to step2. This is repeated until the number of the offspring is N .
- This step is replacement. All chromosomes for the generation are added into global generation. Global

generation will be sorted and top N chromosomes will be selected as final next generation.

- The convergence is measured using standard deviation of fitness values. If the convergence is unsatisfied, set the next generation to current generation and go to step 2; otherwise, stop iteration.
- Sort the last generation by fitness value and select first chromosome as a best solution.

4.1 Chromosome representation

The chromosome representation of our genetic algorithm is order-based representation using linkage learning [9]. Each gene on chromosome represents the index of original matches. The chromosomes are always ordered by linkage value. This mechanism guarantees that the gene including high linkage value would be not separated by crossover with high probability. It is meaningful because fundamental matrix is very sensitive with combination of pairs. In the first generation, to generate a chromosome, L matches should be selected as genes randomly, where L is the length of chromosome. From the second generation, chromosomes are generated by crossover operation.

4.2 Linkage map

A linkage map, which is $N \times N$ matrix, is created to store linkage information. Each value of the matrix is integer value and it represents the linked count. If the i -th gene, which means the index of the i -th match and the j -th gene, is on same chromosome at the same generation, the value of matrix at (i, j) will be increased by 1. Fig. 1 describes how to generate the linkage map in this study.

```

FOR  $i : N$ 
FOR  $i : N$ 
    IF  $i$  is not equal to  $j$ 
         $LM[i][j]++;$ 
    END
END FOR
END FOR
```

Fig. 1. Linkage map generation, where LM stands for a linkage map

The i -th gene on chromosome includes a linkage value which is sum of the i -th row of the linkage map. The linkage value becomes large when the gene survives for long term and exists on many chromosomes.

4.3 Crossover operation

To prevent redundancy of gene value, order crossover operation is used. First, two points p_1 and p_2 less than the length of chromosome are generated randomly. The list of

genes between p_1 and p_2 on the first parent is put into offspring. After then, no redundant genes with the genes of offspring are selected from the second parent.

4.4 Fitness function

Actually, the chromosome represents the list of matches which is selected randomly from original matches. To evaluate the chromosome, fundamental matrix should be generated from the matches of the chromosome. The F matrix is then applied to all matches and the score of each match is calculated by (9):

$$S = x_i^T F x. \quad (9)$$

After then, mean of whole scores is calculated and it is used to calculate fitness value of the chromosome. The final fitness value is calculated such as:

$$\text{Fitness value} = 1 - \text{mean}(S^{1/4}). \quad (10)$$

In our implementation, we calculate the fitness value about only matches satisfying $S < 0.2$ for better performance.

4.5 Genetic algorithm reiteration

Experimentally, we observed that if we use a genetic algorithm one time, the performance is not so good since the wrong matches still alive. In the next iteration of the genetic algorithm, the result of the previous genetic algorithm can be used as input data. As this process runs repeatedly, correct matches will be remained with high probability. The threshold value of the fitness value can be used for stop condition.

5 Implementation and experimentation

We implement the proposed algorithm using OpenCV 2.4.8 library and Microsoft Visual Studio 2010. The parameters used in this study are described in Table 1.

Table 1: Parameters used in this study

Parameters	Value
Length of chromosome	20
Population size	300
Maximum number of generation	30
Threshold value of convergence	0.01
Threshold value of fitness value	0.8

We use two image pairs to evaluate our implementation, as shown in Fig. 2. One image pair is the giraffe doll images

which has many curved surfaces and the other image pair is the eggs images which have three sphere-like objects.

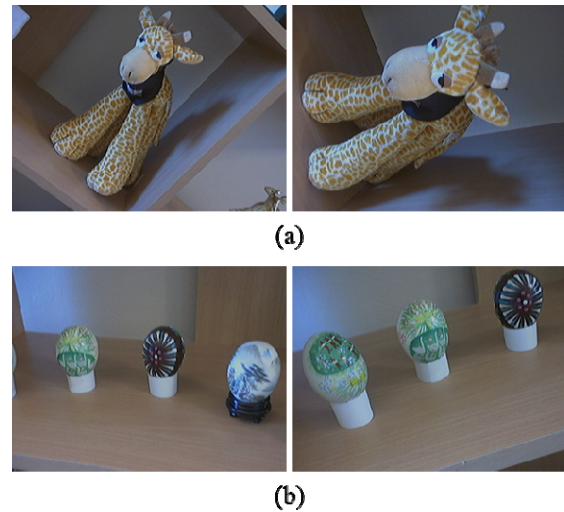


Fig. 2. (a) Giraffe doll images which has many curved surfaces and (b) eggs images which has three sphere-like objects

5.1 Implementation

SIFT features and descriptors are used for matching input data. To generate initial matches, a 2-NN matching method is used. This method matches one feature of query image to two most similar features of train image by comparing descriptor distances. If the distance of first match is shorter than 80% of the distance of the second match, the shorter match is selected as a candidate.

Using the candidate matches, a genetic algorithm runs iteratively and the result is best fundamental matrix which is most fit to candidate matches. Finally, all wrong matches are removed as we described in Section III. Figs. 3 and 4 show results of the proposed images feature matching. Fig. 3 is the result of the 2-NN matching method and Fig. 4 shows selected matches included in the best chromosome.

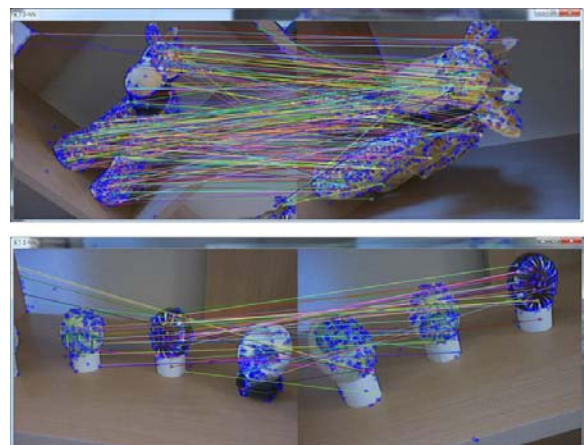


Fig. 3. The result of 2-NN matching

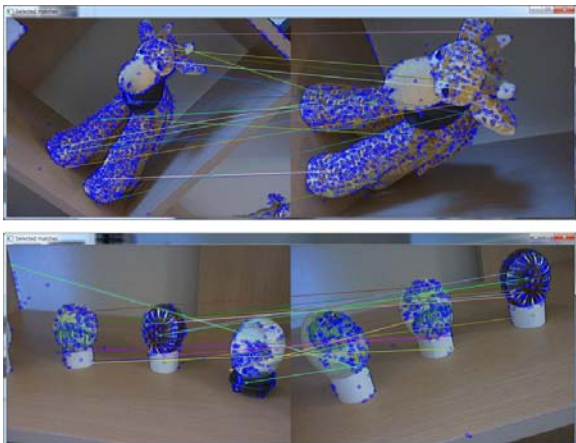


Fig. 4. Selected matches of the best chromosome

Fig. 5 shows the final results of our data sets. As we can see in Fig. 4, some of selected matches are incorrect. In contrast, the result in Fig. 5 is very good since the proposed algorithm extracts fundamental matrix using RANSAC algorithm which is robust to outliers.

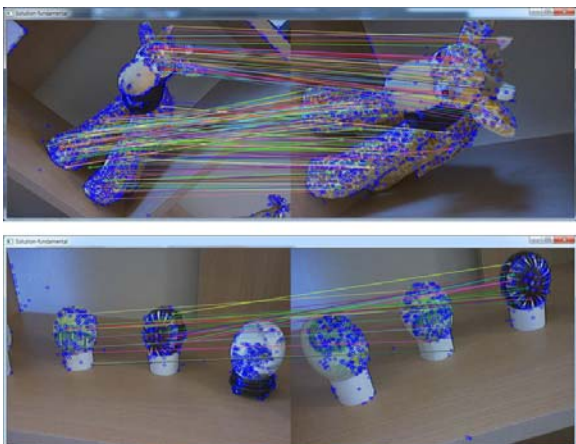


Fig. 5. The results of the proposed method

5.2 Performance evaluation

We compare the proposed method with RANSAC homography matrix estimation, RANSAC fundamental matrix estimation and LMEDS fundamental matrix estimation, which are implemented in OpenCV library. To estimate matrices, we use `findHomography()` and `findFundamentalMat()` function [10].

Fig. 6 shows the results of RANSAC homography matrix estimation. As we assumed, in the case of giraffe images, the result matches are only correctly matched for the plane of legs. In the case of eggs images, there are many incorrect matches since eggs are not plane object.

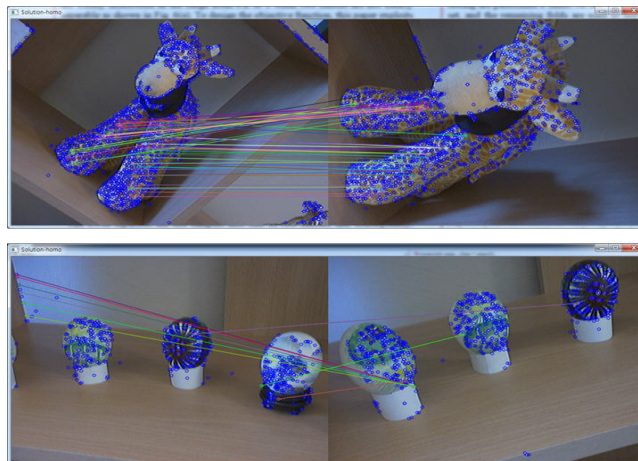


Fig. 6. The results of RANSAC homography matrix estimation

Fig. 7 shows the results of RANSAC and LMEDS fundamental matrix estimations. These methods are very sensitive to the ratio of correct matches and the number of outliers between two different planes. Some incorrect matches are observed in both methods.

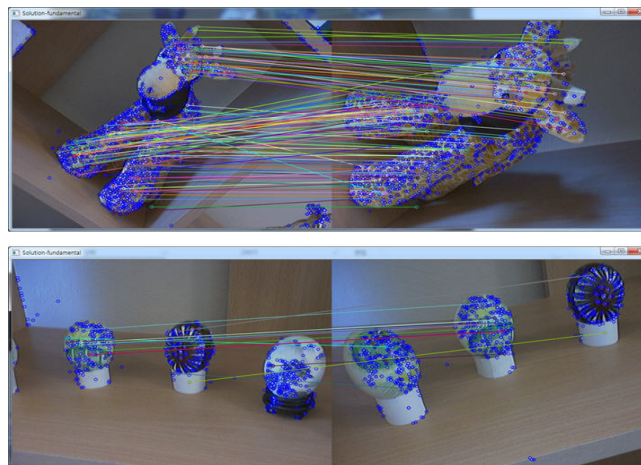


Fig. 7. The results of RANSAC (top) and LMEDS (bottom) fundamental matrix estimation

6 Conclusions

In this paper, we proposed a genetic algorithm-based image feature matching method to solve the correspondence problem among the SIFT-based matches using a genetic algorithm and fundamental matrix. Experimentally, we demonstrated that the proposed method is very efficient to find correct matches of multiple-plane scenes by comparing the performance of the proposed method with other state-of-art methods.

7 References

- [1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [2] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," *International Conference on Computer Vision*, Bombay, Jan. 1998, pp. 754–760.
- [3] W. Wei, H. Jun, and T. Yiping, "Image matching for geomorphic measurement based on SIFT and RANSAC methods," *International Conference on Computer Science and Software Engineering*, Wuhan, Hubei, Dec. 2008, pp. 317–320.
- [4] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," *IEEE International Conference on Computer Vision*, Beijing, Oct. 2005, pp. 1482–1489.
- [5] O. Duchenne, F. Bach, K. Inso, and J. Ponce, "A tensor-based algorithm for high-order graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2383–2395, 2011.
- [6] OpenCV, which is available at <http://www.opencv.org>.
- [7] T. Lindeberg, "Scale-space theory: a basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, vol. 21, no. 2, pp. 224–270, 1994.
- [8] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision." 2nd edition, Cambridge, pp. 222–228, 2001.
- [9] C. Ying-ping, Y. Tian-Li, S. Kumara, and D. E. Goldberg, "A survey of linkage learning techniques in genetic and evolutionary algorithms," *IlliGAL Report*, No. 2007014, Illinois Genetic Algorithms Laboratory, April 2007.
- [10] G. Bradski and A. Kaebler, "Learning OpenCV." 1st edition, O'REILLY, 2008.

Ratios of Eigenvalues for the Dirichlet Laplacian and Hu's Moments

Mohamed A. Khabou¹ and Mohamed B. H. Rhouma²

¹Department of Electrical and Computer Engineering, University of West Florida, Pensacola, FL, USA

²Department of Mathematics, Statistics and Physics, Qatar University, Doha, Qatar

Abstract – We show experimentally that shape features based on the eigenvalues of the Laplacian operator with Dirichlet boundary condition can be accurately approximated using the shape's Hu' moments and vice versa. Simple feedforward neural networks trained using backpropagation algorithm were used to approximate the values of one set of features based on the values of the other set. Three sets of images were used to test the hypothesis: a set of simple computer generated shapes consisting of rectangles, triangles, ellipses, and diamonds; a set of randomly generated convex shapes; and a set of non-convex shapes. The features were typically approximated to within less than 5% of their true values.

Keywords: Laplacian eigenvalues, Hu's Moments, neural networks, shape descriptors, feature mapping.

1 Introduction

In 1951, Polya and Szego [1] published the manuscript "Isoperimetric Inequalities in Mathematical Physics" in which they attempted to establish a few links between *geometric* quantities such as perimeter, area, moment of inertia and *physical* quantities such as torsional rigidity, principal frequency for a few simple shapes. Clearly, geometric quantities can be obtained by simple definite integrals, while physical quantities require solving partial differential equations with specific boundary conditions. The majority of the manuscript by Polya and Szego [1] was dedicated to universal inequalities (valid for all domains) giving bounds of the physical quantities in terms of the geometric ones. Obviously, this stems from a belief that geometry and physics are very much related. This relation was then the subject of the famous question "Can one hear the shape of a drum?" asked by Mark Kac in his famous paper [2]. To hear the shape of a drum is to infer information about its shape (domain) from the sound it makes [3]. Mathematically speaking, given a bounded planar domain Ω (representing the shape of a drum) and the sequence of eigenvalues (representing the sound pitch of the drum) $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots \leq \lambda_k \leq \dots \rightarrow \infty$ of the partial differential equation $\Delta u + \lambda u = 0$ in Ω with appropriate conditions on its boundary $\partial\Omega$, can someone determine the shape of the drum (Ω) based solely on the eigenvalues? Here, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Laplacian operator. The question was answered in the negative in 1992 when

Gordon et al [4] constructed two different shapes that had identical eigenvalues. However, with additional restrictions on the type of drum, one *can* hear its shape [5-9]. In fact, Khabou et al. [10, 11, 12] have shown that features based on the eigenvalues of the Laplacian can be successfully used to represent and classify synthetic and natural images. The feature set they used was defined by

$$F = \left\{ \frac{\lambda_1}{\lambda_2}, \frac{\lambda_1}{\lambda_3}, \dots, \frac{\lambda_1}{\lambda_n} \right\} \quad (1)$$

where, $\lambda_1, \lambda_2, \dots, \lambda_n$ are the first n eigenvalues of the Dirichlet Laplacian and n is the number of features we wish to use for a particular recognition and/or classification problem. These features are translation, rotation, and size invariant; are shown to be tolerant of noise and small deformations in the shape; and have values in the [0 1] range.

Another set of translation, rotation, and size invariant features commonly used in shape representation are Hu's moments [13, 14, 15]. These seven centralized and normalized moments are defined as

$$I_1 = \eta_{20} + \eta_{02} \quad (2)$$

$$I_2 = (\eta_{20} - \eta_{02})^2 + 4 \eta_{11}^2 \quad (3)$$

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (4)$$

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2 \quad (5)$$

$$I_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) [(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (6)$$

$$I_6 = (\eta_{20} - \eta_{02}) [(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \quad (7)$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) [(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (8)$$

where,

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{(1+\frac{i+j}{2})}} \quad (9)$$

$$\mu_{pq} = \int \int_{\Omega} (x - \bar{x})^p (y - \bar{y})^q f(x, y) dA \quad (10)$$

and \bar{x} and \bar{y} indicate the centroid of the shape represented by the function $f(x, y)$.

In this paper, we show numerically that there is a strong relationship between the geometric Hu's moments and the "physical" ratios of the Dirichlet Laplacian eigenvalues. Both of these features were used successfully in recognition problems as they are size, translation and rotation invariant. While there are several incentives for establishing such a link, we will mention here the computational cost for obtaining the ratio of the eigenvalues compared to the straight forward computation of the Hu moments. In fact, comparing the computational cost of the Hu's moments to that of the eigenvalue-based features, we notice that the computational cost of the later grows in a $O(n^2)$ fashion, where n is the number of pixels in a shape, compared to that of the Hu's moments (see Fig 1). It goes without saying that this provides us with a computationally cheap way to approximate the values of the eigenvalue-based features using the Hu's moments.

This paper does not attempt to formalize or explain the link between the geometrical and physical quantities as this is a challenging task that will be the focus of further research.

In the next section we report on the experimental results we achieved with three different data sets and we summarize our findings in the conclusion section.

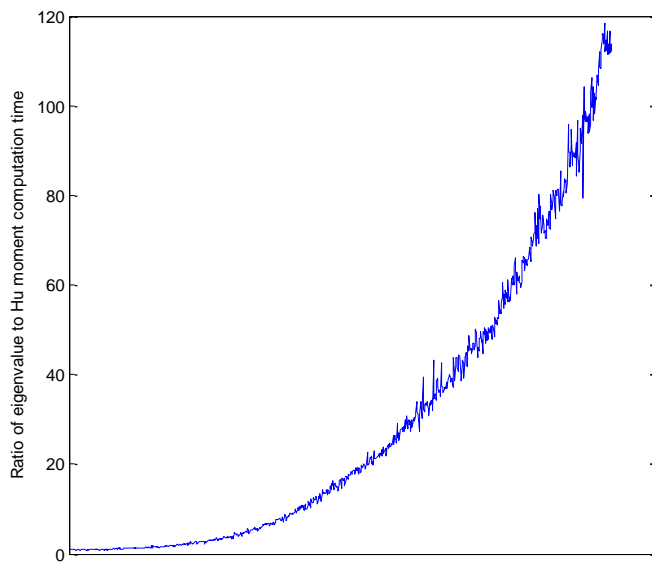


Figure 1. Ratio of computational time of eigenvalue-based features to that of Hu's moments as shape size increases

2 Experimental Results

We conducted three sets of experiments using three different data sets. The first data set consists of 500 computer generated images (100 images per class) of ellipses, rectangles, diamonds, triangles, and squares of different sizes, orientations and aspect ratios (for ellipses, diamonds, and rectangles) (see Fig 2). The second data set

consists of 200 images of random convex shapes (see Fig 3). The third data set consists of 200 images of non-convex shapes that we called the "petal" images since they resemble daisies with different number of petals (see Fig 4). We report on the results of these experiments in the following subsections.

2.1 Estimating Eigenvalue Features using Hu's Moments

In this first experiment we wanted to see if we can estimate the values of the first few eigenvalue-based features of the computer generated shapes using their Hu's moments. The 500 images of the simple shapes were equally split into training and testing subsets. The eigenvalue-based features and the Hu's moments of all shapes were computed. We trained a simple feedforward neural network using the first 4 Hu's moments of the training subset as input and the first 5 eigenvalue-based features as output. The neural net had a single hidden layer composed of 6 hidden neurons. This network structure produced the best results among the other structures we tried. We decided to use only the first 4 Hu's moments for input because we noticed that the remaining moments (I_5-I_7) were all very close to zero for all shapes and hence do not carry much distinctive information. We decided to approximate only the first 5 eigenvalue-based features because, based on our previous work [9, 10, 11], the first "few" features carry the most amount of discriminatory information between shapes. In addition, our goal for this experiment was to show that at least some eigenvalue-based features can be accurately approximated using Hu's moments and hence some sort of a mapping exists between the two. The neural network was trained using the backpropagation algorithm on the training subset and tested using the images in the testing subset. Table 1 summarizes the results.

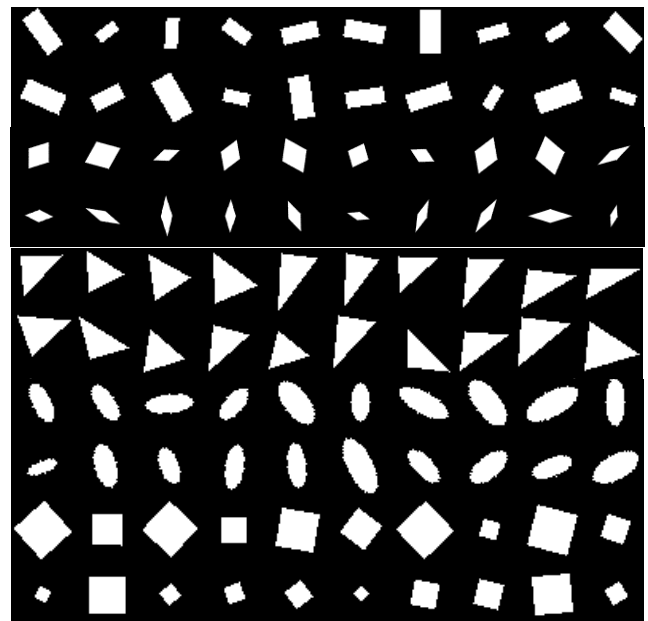


Figure 2. Samples of computer generated images of diamonds, rectangles, triangles, ellipses, and squares.

Table 1. Error analysis on the approximation of the eigenvalue features using Hu's moments

Error Type	Error Magnitude (%)
Average error on λ_1/λ_2	0.0245 (4.8%)
Average error on λ_1/λ_3	0.0117 (3.2%)
Average error on λ_1/λ_4	0.0066 (2.2%)
Average error on λ_1/λ_5	0.0107 (4.6%)
Average error on λ_1/λ_6	0.0087 (4.2%)
Overall average error	0.0125 (3.8%)

As can clearly be seen in Table 1, we were able to approximate the first 5 eigenvalue-based features to within less than 5% of their true value using only the first 4 Hu's moments and a simple non-linear mapping mechanism (neural network). This experiment indicates that there is a relationship between Hu's moments and eigenvalues. What is interesting about this relationship is that the eigenvalue-based features can now be accurately approximated using the much computationally cheaper Hu's moments fed to a simple neural network instead of being computed from scratch. This is especially useful for large shapes where the computational complexity of the eigenvalues is 100 to 1000 folds more than that of the Hu's moments (see Fig 1). The other interesting thing about this relationship is that it relates two sets of features that are completely different in nature and supposedly capture very different aspects of a shape. Obviously, this relationship is yet to be fully understood and mathematically explained—a task the authors are currently undertaking.

2.2 Estimating Hu's Moments using Eigenvalue Features

In this second experiment we wanted to examine the existence of the reverse mapping from eigenvalue features to Hu's moment. For this experiment we used the same shape dataset as the first experiment above. A feedforward neural network with 5 input unit, one hidden layer with 8 hidden units, and 4 output units was trained with the first 5 eigenvalue features as input and the first 4 Hu's moments as output. This structure produced the best results among other structures we tried. Table 2 below summarizes the results.

Table 2. Error analysis on the approximation of the Hu's moments using the eigenvalue features

Error Type	Error Magnitude
Average error on I_1	0.0027
Average error on I_2	0.0016
Average error on I_3	0.0010
Average error on I_4	0.0004
Overall average error	0.0014

As can clearly be seen in Table 2, we were able to approximate the first 4 Hu's moments to within ± 0.0014 of their true values using the eigenvalue features as input and a simple non-linear mapping mechanism. This experiment

shows that a reverse mapping function from the eigenvalue feature space to the Hu's moment space does exist. The nature of this mapping is being studied by the authors. Even though computationally speaking it does not make sense to try to approximate Hu's moments using eigenvalues given that the later are more computationally expensive, the nature of the mapping between the two spaces truly intrigues the authors since, to our knowledge; no one has studied this relationship before.

2.3 Experiment with Random Convex and Non-convex Shapes

In this third experiment we wanted to examine whether the mapping between eigenvalue features to Hu's moments holds for more complex convex and non-convex shapes. For this experiment we created a set of 200 random convex shapes (see samples in Fig 3) and another set of 200 non-convex shapes (see samples in Fig 4). Each set was equally divided into training and testing subsets. The 7 Hu's moments and first 10 eigenvalue features of all shapes were computed and a neural network with 7 inputs (corresponding to the 7 Hu's moments), a hidden layer with 10 hidden units, and 10 outputs (corresponding to the 10 eigenvalue features) was trained on the training subsets and tested with shapes from the testing subsets. This network structure produced the best results among others we tried. This network was able to approximate the eigenvalue features to an average of $\pm 2.61\%$ of their actual values for the *training* subset and $\pm 3.88\%$ for the *testing* set.

This experiment shows that the mapping between Hu's moments and the eigenvalue features still holds for "random" convex and non-convex shapes alike. We plan to scrutinize this result further and try to find a solid mathematical foundation for it. For example, we plotted in figure (Fig 5) the first Hu's moment (I_1) vs. the first ratio of eigenvalues (λ_1/λ_2) for all shapes tried in this paper to see if a pattern emerges. As can be seen from Fig 5, it looks like there is a pattern relating these two measures and some clear boundaries where this relationship exists.

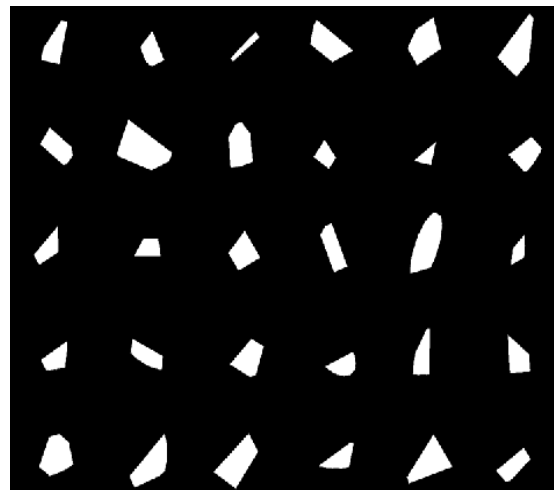


Figure 3. Samples of the random convex images.

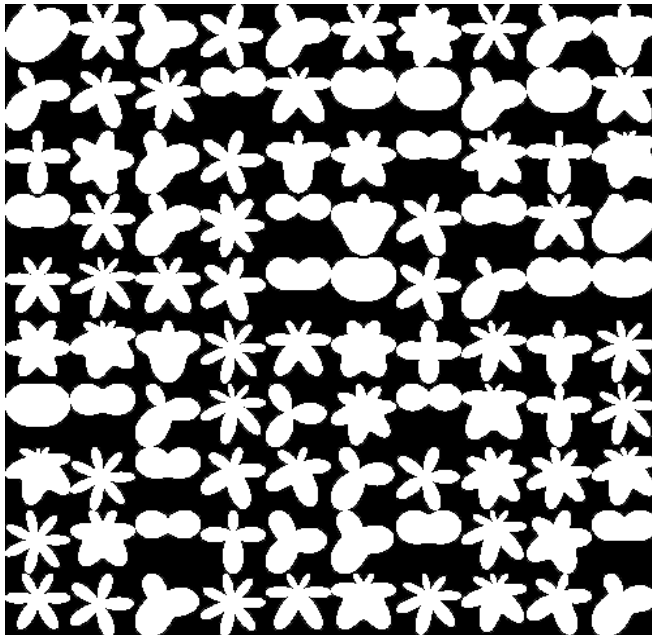


Figure 4. Samples of the “petal” images.

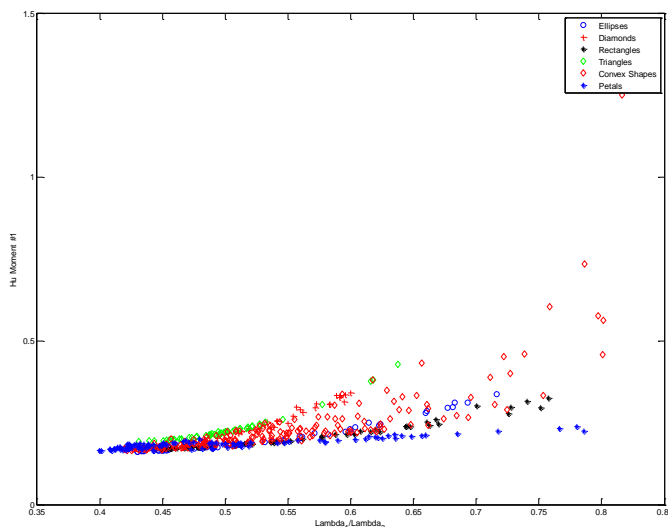


Figure 5. Plot of I_1 vs. λ_1/λ_2 for all shapes in this paper.

3 Conclusion

We showed experimentally that Hu’s moments of a shape can be used to approximate its Laplacian eigenvalue ratios and vice versa. Hu’s moments were used to approximate the eigenvalue-based features of a variety of convex and non-convex shapes to within a $\pm 5\%$ error margin. A simple non-linear mapping tool consisting of a neural network with one hidden layer and very few hidden neurons was used to do the mapping between the two sets of features. The approximation of the eigenvalue features based on the Hu’s moments allows us to reduce the computational complexity of the eigenvalues by a factor of up to 1000—depending on the size of the shape.

The nature of the mapping between the two sets of features requires a closer look to establish a solid theoretical

foundation for it—something the authors plan to pursue as a continuation of this research.

4 Acknowledgement

This work was partially supported by startup grant QUSG-CAS-MPS-12/13-24 from Qatar University in Doha, Qatar. This paper was also partially supported by UREP grant # (14-085-1-011) from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

5 References

- [1] G. Pólya, G. Szego, *Isoperimetric inequalities in mathematical physics* (No. 27). Princeton University Press (1951)
- [2] M. Kac, “Can one hear the shape of a drum?”, *Am. Math. Mon.*, vol. 73, pp. 1–23, (1966)
- [3] M. H. Protter, “Can one hear the shape of a drum? Revisited”, *SIAM Rev.* vol. 29, pp. 185–197, (1987)
- [4] C. Gordon, D. Webb, S. Wolpert, “One cannot hear the shape of a drum”, *Bull. Am. Math. Soc.*, vol. 27, pp. 134–138, (1992)
- [5] M. Zuliani, C. Kenney, S. Bhagavathy, B.S. Manjunath, “Drums and Curve Descriptors”, UCSB Visitation Research Lab Preprint, (2004)
- [6] <http://vision.ece.ucsb.edu/publications/04BMVCMarco.pdf>
- [7] R. Courant, D. Hilbert, *Methods of Mathematical Physics*, second ed., Interscience Publishers, New York, (1965)
- [8] M. Reuter, F. Wolter, N. Peinecke, “Laplace-Beltrami Spectra as Shape DNA of Surfaces and Solids”, *Computer-Aided Design*, vol. 38, pp. 342–366, (2006)
- [9] M. Reuter, F. Wolter, N. Peinecke, “Laplace-spectra as fingerprints for shape matching”, *Proc. of the 2005 ACM symposium on Solid and physical modeling*, pp. 101–106, Cambridge, Massachusetts, (2005)
- [10] M. A. Khabou, L. Hermi, M. B. H. Rhouma, “Shape Recognition Using Eigenvalues of the Dirichlet Laplacian”, *Pattern Recognition*, vol. 40, pp. 141–153, (2007)
- [11] M. B. H. Rhouma, L. Hermi, M. A. Khabou, “Laplacian and Bilaplacian Based Features for Shape Classification”, *Proc. Int’l Conference on Image Processing, Computer Vision, and Pattern Recognition*, pp. 615–619, Las Vegas, NV, (2009)

- [12] M. B. H. Rhouma, M. A. Khabou, L. Hermi, "Shape Recognition Based on Eigenvalues of the Laplacian", chapter in *Advances in Imaging and Electron Physics* vol. 167, pp. 183-252, P. W. Hawks (Ed): Elsevier (2011)
- [13] M. K. Hu, "Visual Pattern Recognition by Moment Invariants", *IRE Trans. Info. Theory*, vol. IT-8, pp. 179-187, (1962)
- [14] J. Flusser, "On the Independence of Rotation Moment Invariants", *Pattern Recognition*, vol. 33, pp. 1405-1410, (2000)
- [15] J. Flusser, T. Suk, "Rotation Moment Invariants for Recognition of Symmetric Objects", *IEEE Trans. Image Proc.*, vol. 15, pp. 3784-3790, (2006)

Distortion Center Estimation Technique Using the FOV Model and 2D Patterns

Dae Hyuck Park
SANE CO., Ltd.
Seoul, Korea
dh.park@sane-auto.com

Jeong Goo Seo
SANE CO., Ltd.
Seoul, Korea
jg.seo@sane-auto.com

Haijung Choi
SANE CO., Ltd.
Seoul, Korea
hj.choi@sane-auto.com

Eui Sun Kang
Soongsil University
Seoul, Korea
kanges@naver.com

Abstract— This paper proposes a method for estimating a distortion center to correct the radial distortion that occurs when capturing images with a wide-angle fish-eye lens. In the field-of-view (FOV) distortion correction model, the error of the distortion center and center of the image increases because it does not estimate the distortion center of a lens separately. This drawback deteriorates the accuracy of distortion correction. Thus, this paper proposes a distortion correction method using the FOV model and 2D patterns in order to increase the accuracy of the distortion center estimation of a wide-angle lens. To achieve this goal, a distortion curve generated from the FOV model is compared with a straight line, thereby setting the position of the minimum difference between the curve and the straight line as the distortion center. Through this method, the accuracy of the estimation of the center of the distortion that occurs owing the error in the alignment of center points of a lens and the imaging sensor can be improved. This was also verified through experiments.

Index Distortion center estimation, FOV model, 2D pattern, around view monitor, 190° wide-angle camera, AVM(Around View Monitor), SVM(Surround View Monitor)

I. INTRODUCTION

A driver assistance system that aims to reduce the traffic accident rate analyzes data collected from various sensors installed in a vehicle and provides the analysis result to the driver. One such technology is the Around View Monitor (AVM) system in which the data received by three or more camera devices mounted on the vehicle are used to provide the driver with surrounding image information to prevent collision while driving or parking a car.

In general, a passenger car is equipped with four cameras with wide-angle lenses mounted on the front, rear, left, and right sides of a vehicle to capture the maximum field of horizontal and vertical images from the surroundings. A wide-angle lens, which can capture a picture with a wide angle greater than 120° with a short focal length, generates radial distortion by which a ray of light that enters the lens farther from its center is more curved than a ray of light that enters the lens closer to its center due to the effect of a curved lens. This phenomenon often occurs in fish-eye lenses used in AVM

systems, and such distortion is more severe near the edge of the image than at the center.

In order to correct the distortion of a fish-eye lens, two methods have been used: approximation of distortion functions into polynomial forms and the field of view (FOV) model, which is a geographical model based on non-linear distortion characteristics. In the polynomial distortion model, computational complexity increases as the order of polynomials increases and the difficulty of application increases as the view angle of the fish-eye lens becomes larger. However, the FOV model is more efficient than the polynomial distortion model because its design is based on the non-linear distortion of fish-eye lenses, although it has a problem of adding distortion when there is an error in the location of the distortion center.

To solve this problem, this paper proposes a method to estimate a distortion center, which serves to accurately correct the distortion that occurs in the camera image signals from an ultra-wide viewing angle greater than 190°. In particular, this paper proposes a method to find the center point of distortion that can minimize distortion using a lattice-patterned 2D plane and the FOV distortion model to correct distortion.

This remainder of this paper is organized as follows. In Section II, previous studies related to distortion correction are described, whereas the proposed distortion center estimation method using 2D planes is described in Section III. In Section IV, the experiment environment and results are discussed, and the conclusion is presented in Section V.

II. RELATED RESEARCH

A complex calculation is required to transform 3D images from cameras into 2D planar images that can be processed by computers. To reduce such computational complexity, a pinhole camera model is used. A pinhole camera, which was used to locate the center point of an image, converts 3D information into 2D planar pixel units based on the optical image received through a small hole. In order to convert 3D spatial coordinates into 2D image coordinates, external parameters such as the installation height and direction of the

camera, and internal parameters such as the focal length and center point of camera are required. The focal length in the internal parameter refers to a distance between the focal point and the image sensor CCD (Charge Coupled Device) and CMOS (Complementary Metal Oxide Semiconductor), which is represented by

$$x_{screen} = f_x \left(\frac{x}{z} \right) + C_x \quad y_{screen} = f_y \left(\frac{y}{z} \right) + C_y \quad (1)$$

Here, x_{screen} and y_{screen} refer to the coordinates on the 2D plane, whereas f is the distance between the focal point and the image plane. In addition, Z is the distance between the object and the focal point, whereas C is the displacement of the coordinate center in the projection plane. Using Eq. (1), the location where the image appears on the 2D plane can be calculated. However, because only a small amount of light passes through a pinhole camera, a long exposure time is required to create an image. In order to collect a large amount of light, a curved lens is used, thereby obtaining images by collecting the curved light. When the obtained image is projected onto a 2D plane, a problem of image distortion can occur due to the characteristics of the lens. The distortion by a lens can be divided into two types: radial distortion, which is generated more severely in an area farther from the center, and tangential distortion, which creates an elliptical distortion distribution.

Fish-eye lenses are typically attached to vehicles, and they can create radial distortion. To resolve the radial distortion problem, three methods can be employed: the method of using the center point of distortion, distortion parameters, and internal parameters; the method of performing polynomial distortion iteratively to transform the distorted curves caused by radial distortion into straight lines; and the method of using image information only. Heikkila [1] proposed a method of finding the center of distortion and internal parameters by using chessboard-like images, in which a method to find parameters that can integrate distortion correction and a camera calibration process was introduced. However, this method has a drawback that could increase iterative calculation complexity when the distortion is excessively severe, although it can be efficient when distortion is moderately severe.

In [2], three orthogonal planar patterns were introduced to cover an entire 180° image with a specific pattern of asymmetrical distortion. This method locates a vanishing point where distorted curves converge to a single point and then defines that as the center point of distortion, thereby performing distortion restoration. This method does not depend on distortion models of parameters due to the special structure of the apparatus, and therefore, it has an advantage in that it can be applied to various lenses, although it is not appropriate for a case where radial asymmetric distortion is generated.

In [3], a center radius was defined using the center of a sphere and the position of a single distorted point. A corrected position value was then used to restore the distorted curves into straight lines. Then, new radii at all positions in the image were obtained to be applied in the FOV model. This method can be

used for real-time processing because it uses low-order polynomials to change curves into straight lines, although it has a weakness in terms of accuracy compared to other existing methods.

In [4], distortion parameters based on the assumption that lines were straight prior to distortion were found using the characteristics of the pinhole camera model in order to minimize the curvature of the lines, and then, they were applied to the tangential distortion correction method.

In [5], the longest curved lines were extracted and removed from 2D images, and then the correction of the remaining curves was performed. However, this method has several drawbacks in that curve detection is performed slowly whereas the removal of the curves may not be carried out accurately owing to the use of a single fixed threshold value for curve removal. To overcome these drawbacks, a method was proposed in [6] to detect and remove the lines quickly using Hough transformation.

The FOV model, which is based on non-linear lens distortion characteristics, corrects distortion under the assumption that the center of an image and the center of lens distortion are the same. However, in the case of cameras using lenses with special functions and a number of layers, an error of the center point might occur during the manufacturing process, thereby creating a fine center point error in the projection onto the 2D plane. Therefore, not only can distortion correction not be performed correctly, but also additional distortion of the image can occur after distortion correction. To solve this problem, a method was proposed in [8] to correct distortion after estimating a center by selecting three straight lines in the image plane to decrease computation complexity.

In addition, a method was presented in [6] to estimate the center of distortion by extracting a curve from the projected 2D image and modifying it into an undistorted straight line, thereby finding the direction in which the center of the extracted curve is changed. However, its accuracy varies depending on the number of detected lines and the existence of lines around the distortion center.

This paper proposes an accurate distortion correction method through the estimation of the distortion center of a lens using the FOV model and 2D patterns in order to correct radial distortion.

III. DISTORTION CENTER ESTIMATION METHOD USING FOV MODEL AND 2D PATTERNS

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

A. Distortion correction method considering distortion center estimation

The process of distortion correction considering the distortion center is shown in Fig. 1. In this paper, a 2D planar image of a chessboard pattern was used first to correct distortion of an image from a fish-eye lens quickly. Using the distorted chessboard pattern, a distortion coefficient was estimated by applying the FOV model.

The center of distortion was found using the estimated distortion coefficient and distorted curve component. Based on the center of distortion (C_x, C_y), a lookup table LUT (is produced, which represents the relationship between distorted location and 2D planar location using the FOV model. Distortion can be corrected by applying the LUT produced offline to real images.

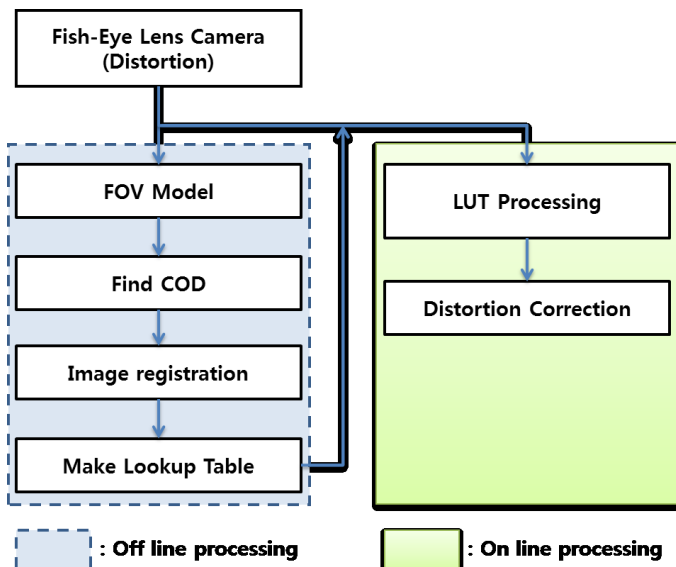


Fig. 1 Distortion correction algorithm of 2D planar pattern

B. FOV distortion model

The FOV model [3,4] calculates a location value in the image over the plane through the coordinate in which distortion occurs when radial distortion is detected once a video image is acquired by a fish-eye lens. The following figure shows the FOV model.

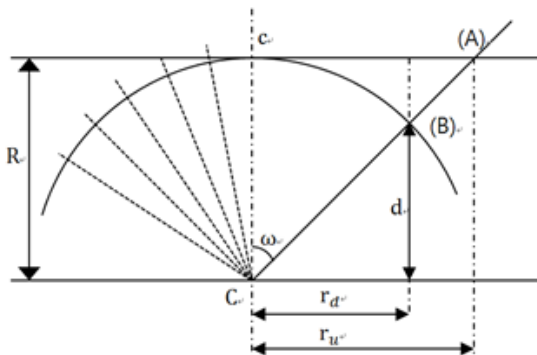


Fig. 2 FOV model

The distance from the center of a sphere to one point of the planar image is r_u , whereas r_d is the distance to the distorted location projected onto the sphere plane. Since the FOV model is based on the optical model, it is derived by trigonometric functions with regard to the angle ω . Once a point (A) is moved to a point (B) due to radial distortion projected onto the plane of the sphere and acquired by a fish-eye lens, it is projected onto the image sensor, thereby producing radial distortion. The FOV model can compute r_d and r_u using the distortion functions and their inverse functions.

$$r_d = \frac{1}{2\omega} \arctan(2r_u \tan \omega) \tag{2}$$

$$r_u = \frac{\tan(r_d \tan \omega)}{2 \tan \omega} \tag{3}$$

The above equations are rearranged using radius R as

$$\frac{R \times r_d}{\sqrt{R^2 - r_d^2}} \tag{4}$$

C. Distortion coefficient estimation of the FOV model

According to the principle of the camera model [7], a straight line in 3D should be a straight line after being projected onto a 2D surface by a camera if there is no distortion. This means that the larger the difference between an actual straight line component and a projected straight line, the larger the distortion is. Conversely, the less the difference is, the less the distortion is. The degree of camera distortion can be verified by the distortion coefficient ω of the FOV model. In the FOV model, $q_{\omega i}$ is a point of restoration of distorted point p_i by the distortion coefficient ω . The distortion parameter estimation from [10] expresses a relationship of $D^{-1}(\omega, p_i)$ with regard to the distortion coefficient, which can solve the linear equation of the least squares distance when the estimation is applied to the algorithms from p_1 to p_n .

By solving the equation of the i and j functions where the error function $E_{ij}(\omega)$ is minimized with regard to the distortion coefficient ω , the distortion coefficient ω can be estimated. Using this method, distortion correction can be performed by using the distortion correction coefficient with regard to the camera distortion center.

$$q_{\omega i} = [x_{\omega i} \ y_{\omega i}]^T = D^{-1}(\omega, p_i) \tag{5}$$

$$E(\omega) = \sum_{i=1}^n \|y_{\omega i} - L(x_{\omega i})\|^2 \tag{6}$$

$$\arg \min_{\omega} \sum_{i=1}^n \sum_{j=1}^m E_{ij}(\omega) \tag{7}$$

Using the distortion coefficient estimation method, once a chessboard pattern is captured, distortion correction can be executed quickly using only the distortion coefficient. Nonetheless, a fine error in the components of the row and column of the chessboard was discovered through experiment. In the case of an ultra-wide-angle camera, the calibration error becomes larger near the edge of the image than at the center owing to the characteristics of the lens. Thus, distortion correction using the FOV model is appropriate.

TABLE I. MEASUREMENT OF DISTORTION CENTER ESTIMATION RESULT IN THE AVM CAMERA

	Cod_x	Cod_y
CAM #1	10	-9
CAM #2	10	0
CAM #3	4	3
CAM #4	-4	7

D. Distortion center estimation method using 2D patterns

The distortion center estimation method using 2D patterns is proposed to overcome the limitation of the distortion correction method using the distortion coefficient ω of the FOV model. In this paper, a chessboard pattern was shot to estimate the center of distortion followed by projecting it onto an actual straight line in the image data of corrected distortion, thereby estimating the center of distortion using the value of the distance difference between the straight lines.

In order to estimate the center of distortion, first the distortion in the 2D patterns was corrected using the distortion coefficient of the FOV model. Then, a certain range surrounding the center of distortion, which was determined while applying the FOV model, was set to the detection window of the center of distortion.

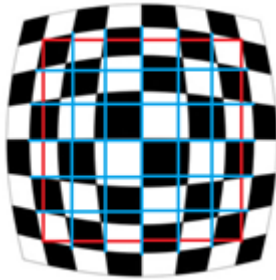


Fig. 3 Example of the detection window of the distortion center and straight-line components.

Next, straight lines are generated by following the detected points using the method described in Ref. [11] in the detection window of the center of distortion. Figure 3 shows the straight lines used to compute the distortion error and detection window in which the distortion center is expected. In the chessboard pattern, where a number of crossing points of

straight lines can be found; crossing points of $M \times N$ representing every corner point are found, and then vertical M straight lines and horizontal N straight lines using the main outer points are generated based on the outermost points, thereby determining whether crossing points are present in the center within the straight lines. The smaller the vertical and horizontal distortions are, the closer the point is to the center of distortion, which also means it is closer to a straight line. Thus, a center of distortion can be estimated by computing the error between the straight line and the distortion. To estimate the center of distortion, distances in the row and column directions are added, and an equation for the straight line that is closest to the center of distortion and C_x, C_y is solved. Therefore, the distances to the distortion points (p_{ij, c_x, c_y}), where there are error points with the straight line (L_{i, c_x, c_y}) are computed.

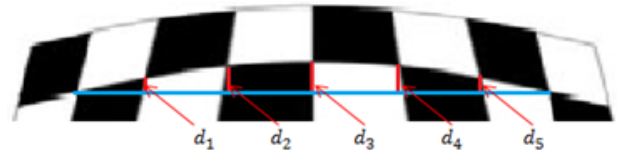


Fig. 4 Distortion error

Figure 4 shows an example of errors between the distortion and a straight line, which represents distances d_n between the blue-colored straight line and the distortion in the 2D chessboard pattern. The points detected in the figure are used to determine the degree of distortion. That is, the distances between the crossing points and the straight line are computed in the vertical and horizontal directions, thereby summing the distances of the inner points ($d_1 + \dots + d_n$) to calculate the distortion error.

$$E_{ij}(c_x, c_y) = \|L_{i, c_x, c_y} - p_{ij, c_x, c_y}\|^2$$

$$E_{ji}(c_x, c_y) = \|L_{j, c_x, c_y} - p_{ji, c_x, c_y}\|^2 \quad (8)$$

$$\arg \min_{c_x, c_y} \left(\sum_{i=1}^n \sum_{j=1}^m E_{ij}(c_x, c_y) + \sum_{j=1}^m \sum_{i=1}^n E_{ji}(c_x, c_y) \right) \quad (9)$$

In other words, it finds the minimum sum-of-squares difference between the actual straight line component and the projected line component. By using this function for distance computation, the minimum distance to the straight line component of row (m) and column (n) is calculated to estimate the distortion center (c_x, c_y).

The distortion center estimation method using 2D patterns performs precise distortion correction by finding the minimum distortion distance. Precise distortion correction can be achieved by applying the least distortion distance estimation method to the FOV distortion correction model using 2D patterns. An LUT is produced with regard to the distortion

locations on the 2D plane using the estimated distortion center, thereby being applied to the actual image.

IV. EXPERIMENT AND EVALUATION

In order to verify the result of the distortion correction algorithm, the following distortion correction experiment device was developed. The experiment environment was configured using notebook computers, USB cameras, wide angle lenses, and chessboard patterns as the hardware configuration. The target image was a 2D black and white chessboard pattern with 6 horizontal rows and 11 vertical columns. One square of the chessboard was $163_{pixel} \times 163_{pixel}$ while the actual size was 45 mm square.

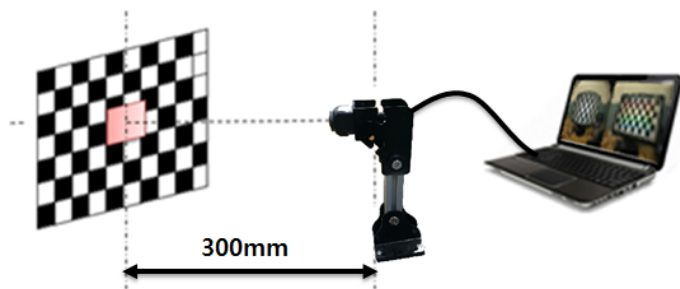


Fig. 5 Experiment environment for distortion correction using wide angle lens camera

As shown in Fig. 5, the experimental environment was constructed to perform real-time processing for the proposed algorithm, in which a 2D chessboard was positioned at the center, and a real-time camera image taken with a wide-angle lens was sent to the personal computer for distortion correction.

Once the accurate distortion center (C) is set in the FOV model and the distortion is corrected, all straight lines can be restored almost completely as shown in Fig. 6. This result is obtained because the distortion center was set repeatedly in the distortion correction process.

However, additional distortion was generated as shown in Fig. 7 when distortion correction was performed again assuming that the distortion center was the center point of the image where there was an error from the lens distortion center. The analysis result of the distortion center showed that when a fine error of -30 pixels in the X direction and $+30$ pixels in the Y direction was applied to the same image, a phenomenon representing a fine curve with a specific directivity was found. This means that more severe distortion was found in proportion to the distortion center. The reason for this fine center error while projecting it onto the 2D plane is due to a mismatch of the center point with regard to the optical axis occurred during the camera manufacturing process in the case of cameras with a number of layered lenses. Therefore, while applying the FOV model, which does not estimate the center of distortion separately, there is a problem of deterioration of distortion correction accuracy as the error of the distortion center of the lens and the center point of the image becomes larger.

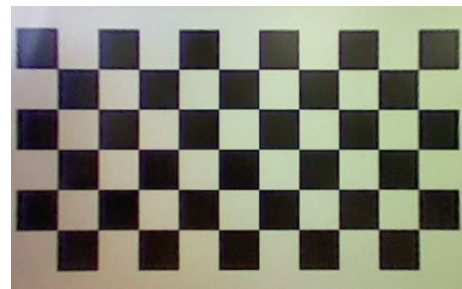


Fig. 6 Distortion correction using the FOV model.

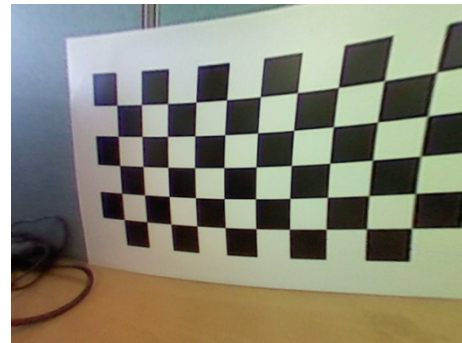


Fig. 7 Result of the distortion center error (X axis: -30 pixels, Y axis: $+30$ pixels)

The experiment to determine the error of the measurement angle using the Zhang algorithm [12], which is regarded as a representative method of distortion correction, produced four measurement results as shown in Table 2. In addition, it was verified that deflection of the lens center was discovered in the X direction, to the right of the Y axis, and in upper portion of the image.

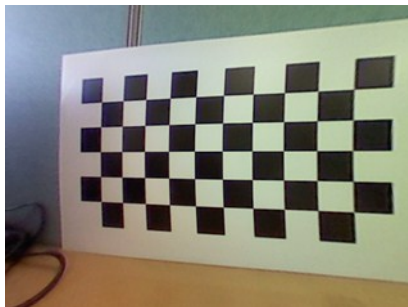
TABLE II. ESTIMATION RESULT OF DISTORTION CENTER USING THE ZHANG ALGORITHM

	Cod_x	Cod_y
First measurement	12.1	-1.13
Second measurement	12.03	-1.01
Third measurement	6.96	-1.87
Fourth measurement	10.43	-0.78

Accordingly, this paper solved the problems that straight lines were expressed as curves due to no estimation of the distortion center in the FOV model and deflected representation by means of precise correction of the distortion center. Figures 8 and 9 show the correction result after the distortion center in the FOV model was estimated in the horizontal and vertical directions.



(a) Estimation of the distortion center in the vertical direction



(b) Estimation of the distortion center in the horizontal direction

Fig. 8 Estimation of the distortion center in the vertical and horizontal directions using the FOV model.

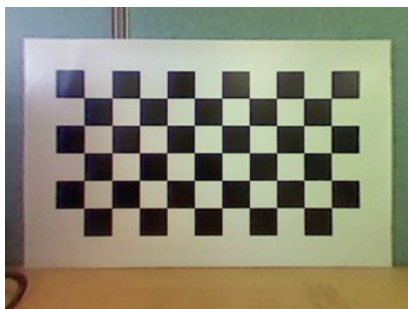


Fig. 9 Correction result using the distortion center estimation method.

In the case of incorrect distortion correction due to the displaced distortion center, the image was corrected only in either the vertical or horizontal direction as shown in Figs. 8(a) and (b), respectively. This phenomenon occurred because of the incorrect designation of the distortion center axis for the distortion correction. To minimize this error, the precise center axis can be found using the distortion center estimation method by means of 2D patterns such as a chessboard pattern. This resulted in minimizing the distortion correction error that might occur owing to instrument error while configuring the experiment or camera mounting.

TABLE III. RESULT OF THE PROPOSED DISTORTION ESTIMATION METHOD

	Cod_x	Cod_y
First measurement	10	-11

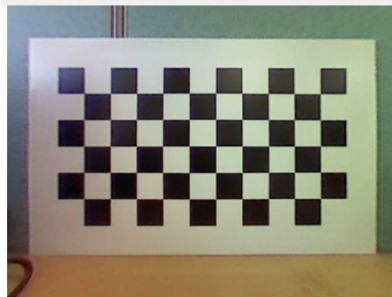
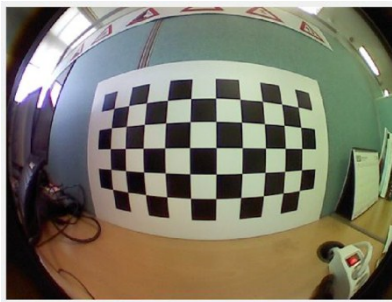


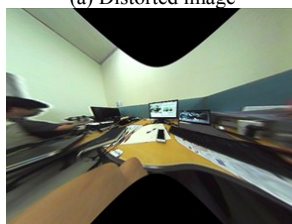
Fig. 10 Result of the distortion center correction using 2D planar patterns.



(a) Distorted image



(b) Distorted correction result



(c) Distorted correction result



(d) Distorted correction result

Fig. 11 Experiment result using actual image

Figure 11 shows the distortion correction results as image data was received in real time from the 190° wide-angle camera. As shown in Fig. 11(a), distorted correction can be found in the distorted monitor image in the horizontal and vertical directions. Figure 11(b) shows the distorted result image when the original image was corrected while (c) and (d) show the full unwrapped images of (b). If images collected from a wide-angle lens are corrected in a distorted manner, they are transformed into images of infinite size so that cropping at an appropriate scale should be used to produce images for monitoring and composition.



(a) Image composition prior to application of the distortion center estimation value



(b) Image composition after application of the distortion center estimation value

Fig. 12 Comparison between image compositions before and after distortion center estimation.

As shown in Fig. 12, there was a significant difference in the image interface matching before and after the distortion center estimation application. The distortion center estimation can be more influential on the composition of a number of camera inputs than on a single camera input.

V. CONCLUSION

In recent years, vehicles have used image processing technologies increasingly while black-box systems for vehicles and AVM systems have been widely used as embedded systems for information recording, parking assistance, safe driving, and the prevention of traffic accidents. Such systems record all details of the surrounding areas using an ultra-wide-angle lens (190°) with high resolution. In order to use such images, distortion correction is necessary to enable users to monitor these images.

This paper proposed a distortion correction method for a camera model in which the distortion center correction method was optimized using the FOV model, and distortion center was found using 2D patterns. Through the proposed method, complete distortion correction can be achieved in the vertical and horizontal directions so that images without size and pattern distortion can be used for image processing, image recognition, and monitoring by users. In general, cameras for AVM systems use high-precision center point estimation. Our algorithm can produce high quality and minimum error of

AVM top views from images acquired from inexpensive cameras.

ACKNOWLEDGEMENTS

This work was supported by the Technology Innovation Program (or Technology Innovation Program, "0043358, Information Composition and Recognition System for surrounding images possible for top view and panorama view of resolving power less than 10cm) funded By the Ministry of Trade, industry & Energy (MI, Korea)" and "R&D Infrastructure for Green Electric Vehicle (RE-EV) through the Ministry of Trade Industry & Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT)".

REFERENCES

- [1] J. Heikkila, "Geometric camera calibration using circular control points", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No. 19, pp.1066-1077, Oct. 2000.
- [2] W. Kim, Y. K Baik, and K. M Lee, "A Parameter-free Radial Distortion Correction of Wide Angle Lenses using Distorted Vanishing Points", 14th Japan-Korea Joint Workshop on Frontiers of Computer Vision (FCV), (1), pp.53-58., 2008.
- [3] Y. K. Jo, Barrel Distortion Compensation of Fisheye Lens for Automotive Omnidirectional Camera Module System, Dongguk University. 2010.
- [4] F. Devernay and O. Faugeras, "Straight lines have to be straight-automatic calibration and removal of distortion from scenes of structured environments", *Mach. Vision and Appl*, Vol.13, No. 1, pp.14-24, Aug. 2001.
- [5] T. Thorsten, B. Hellward, "Automatic line-based estimation of radial lens distortion", *Integrated Computer-Aided Engineering*, Vol.12, No. 2, pp.177-190, 2005.
- [6] B. K. Kim, S. W. Chung, M. K. Song, and W. J. Song, "Correcting Radial Lens Distortion with Advanced Outlier Elimination," *IEEE International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 1693-1699, Nov. 2010.
- [7] S. Shah, J. Aggarwal, "Intrinsic parameter calibration procedure for a (higher distortion) fisheye lens camera with distortion model and accuracy estimation", *Pattern Recognition*, Vol.29, No.11, pp.1775-1788, 1996.
- [8] Aiqi Wang, Tianshuang Qiu, Longtan Shao, "A Simple Method of Radial Distortion Correction with Centre of Distortion Estimation", *Journal of Mathematical Imaging and Vision*, Vol.35 No.3, pp.165-172, November 2009.
- [9] Tat-wa Chao, "Wide-scoped Top-view Monitoring and Image-based Parking Guiding", Master's Thesis, 2009
- [10] Tat-wa Chao, "Wide-scoped Top-view Monitoring and Image-based Parking Guiding", Master's Thesis, 2009
- [11] C. Harris and M. J. Stephens, "A combined corner and edge detector," *In Alvey Vision Conference*, pp.147-152, 1988.
- [12] Z. Zhang, "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.11, pages 1330-1334, 2000.

Classification of Protein Crystallization Trial Images using Geometric Features

M. Sigdel¹, M. S. Sigdel¹, İ. Dinç¹, S. Dinç¹, M. L. Pusey², and R. S. Aygün¹

¹Computer Science Department, The University of Alabama in Huntsville, Huntsville, Alabama, United States

²iXpressGenes, Inc., 601 Genome Way, Huntsville, Alabama, United States

Abstract—*In this paper, we describe our method for classification of protein crystallization trial images using geometric features. The objective is to automatically categorize a protein crystal according to the presence of protein crystal types in the images. We consider only the images consisting of protein crystals for the classification. The images are classified into 4 categories- needles, small crystals, large crystals and other crystals. Image classification consists of two main steps - image feature extraction and applying decision tree classifier. Our feature extraction includes application of canny edge detection, extraction of edge related features from the edge image, and extraction of blob related features from multiple thresholding techniques. We performed our experiments on 212 expert labeled images and tested our results using 10-fold cross validation. Our results indicate that the proposed classification technique produces a reasonable classification performance. The overall accuracy of the classification is 75%.*

Keywords: crystallization, edge detection, blob features

1. Introduction

Protein crystallization is the process for formation of protein crystals. Protein crystallization is a rare process and requires thousands of trials for successful crystallization [1]. The objective of crystallization trials is to determine suitable conditions for protein crystallization and produce protein crystals suitable for X-ray diffraction.

High throughput systems have been developed in recent years trying to identify the best conditions to crystallize proteins [1]. Imaging techniques are used to monitor the progress of crystallization. The crystallization trials are scanned periodically to determine the state change or the possibility of forming crystals. With large number of images being captured, it is necessary to have a reliable classification system to distinguish the crystallization states each image belongs to. The fundamental aim is to discard the unsuccessful trials, identify the successful trials, and possibly identify the trials which could be optimized.

Many research studies have been done to distinguish the protein images as non-crystal (does not contain crystal) or crystal (has crystal). For example, Cumba et al. (2003)[2], Cumba et al. (2005) [3], Berry et al. (2006) [4], Pan et al. (2006) [5] and Po and Laine (2008) [6] have described

the classification of crystallization trials into non-crystal or crystal categories. In our previous work [7], we described classification of crystallization images into three categories (non-crystals, likely-leads and crystals). Saitoh et al. (2006) [8] proposed crystallization trials classification into five categories (clear drop, creamy precipitate, granulated precipitate, amorphous state precipitate, and crystal). Spraggon et al. (2002) [9] have described classification of the crystallization imagery into 6 different categories (experimental mistake, clear drop, homogeneous precipitant, inhomogeneous precipitant, microcrystals, and crystals). Likewise, Cumba et al. (2010) [10] classified into 6 basic categories (phase separation, precipitate, skin effect, crystal, junk, and unsure).

Not all protein crystals are suitable for X-ray diffraction. The main interest for crystallographers is the formation of large 3D crystals. Other crystal structures are also important as the crystallization conditions can be optimized to get better crystals. Therefore, it is necessary to have a reliable system that distinguishes between different types of crystals according to the shapes and sizes. In the previous studies, classification of the different types of crystals has not been the main focus.

Various classification techniques have been proposed for the classification of protein crystallization trials. Classification algorithms such as support vector machines (SVMs), decision trees, neural networks, boosting, and random forest have been used [7]. Alternatively, combination of multiple classifiers has also been studied in the literature [8]. The recent study by Hung et al. (2014) [11] have proposed protein crystallization image classification using elastic net.

In terms of the feature extraction, a variety of image processing techniques have been proposed. Research studies Cumba et al. (2003) [2], Saitoh et al. (2004)[12] and Zhu et al. (2004) [13] used a combination of geometric and texture features as the input to their classifier. Saitoh et al. (2006) [8] used global texture features as well as features from local parts in the image and features from differential images. Cumba et al. (2010) [10] extracted several features such as basic statistics, energy, Euler numbers, Radon-Laplacian features, Sobel-edge features, microcrystal features, and GLCM features to obtain a large feature vector. Increasing the number of features may not necessarily improve the accuracy. Moreover, it may slow down the classification process.

This study describes our technique for protein crystallization image classification. Our focus is on classifying crystallization trial images according to the types of protein crystals present in the images. Our feature extraction includes edge related features from canny edge image and extracting blob related features from multiple thresholding techniques. The images are classified into 4 categories- needle crystals, small crystals, large crystals and other crystals. Image classification consists of two main steps - image feature extraction and applying decision tree for the classification. We are able to achieve a reasonable classification performance.

This paper is arranged as follows. The following section describes the image categories for the classification problem considered in this paper. Section 3 provides the image processing and feature extraction steps used in our research. Experimental results and discussion are provided in Section 4. The last section concludes the paper with future work.

2. Image Categories

The simplest classification of the crystallization trials distinguishes between the non-crystals (trial images not containing crystals) and crystals (images having crystals). In this study, we are interested in developing a system to classify different crystal types. We consider four image categories (Needle crystals, Small crystals, Large crystals, Other Crystals) for protein crystallization images consisting crystals. Description of each of these categories is provided next.

Needle Crystals - Needle crystals have pointed edges and look like needles. These crystals can appear alone or as a cluster in the images. The overlapping of multiple needle crystals on top of each other makes it difficult to get the correct crystal structure for these images. Fig. 1[a-c] show some sample images under this category.

Small Crystals - This category contains small sized crystals. These crystals can have 2-dimensional or 3-dimensional shapes. These crystals can also appear alone or as a cluster in the images. Because of their small size, it is difficult to visualize the geometric shapes expected in crystals. Besides, the crystals may be blurred because of focusing problems. Fig. 1[d-f] provide some sample images under this category.

Large Crystals - This category includes images with large crystals with quadrangle (2-dimensional or 3-dimensional) shapes. Depending on the orientation of protein crystals in the solution, more than one surface may be visible in some images. Fig. 1[g-i] show some sample images under this category.

Other Crystals - The images in this category may be combination of needles, plates, and other types of crystals. We can observe high intensity regions without proper geometric shapes expected in a crystal. This can be due to focusing problems. Some representative images are shown in Fig 1[j-l].

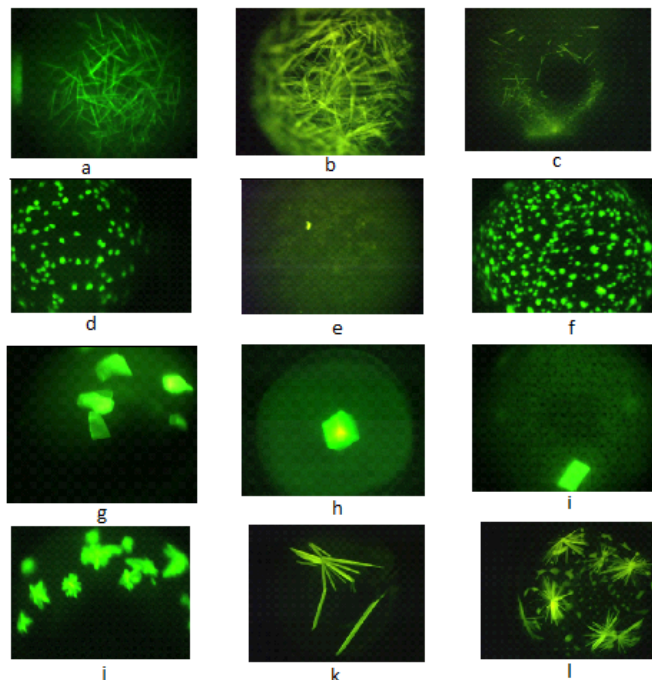


Fig. 1: Sample protein crystallization images: [a-c] Needle Crystals [d-f] Small Crystals [g-i] Large Crystals [j-l] Other Crystals

3. Feature Extraction

The images of crystallization trials are collected using CrystalX2 software from iXpressGenes Inc. Protein solutions are trace fluorescently labeled and the images are collected with green light as the excitation source. As such, the crystals are expected to be highlighted (high intensity) in the image. This can simplify further image processing as the desired objects (crystals) become distinct.

The distinguishing characteristics of protein crystals are the presence of straight lines and quadrangular shapes. Therefore, we focus on extracting geometric features of the objects (or regions) in the image. Fig. 2 shows the components for image pre-processing and feature extraction of our system. Firstly, we down-sample the image and generate binary images using two thresholding techniques. Next, we apply image segmentation and extract features related to the blobs from these binary images. Similarly, we apply canny edge detection and link the edges to get separated segments (graphs) in the image. We then find features related to the segments and the edges. Details of our image processing and feature extraction technique is provided next.

3.1 Image downsampling

A high resolution image may keep unnecessary details and increases the computation time significantly. Therefore, we down-sample the images before further processing. In our

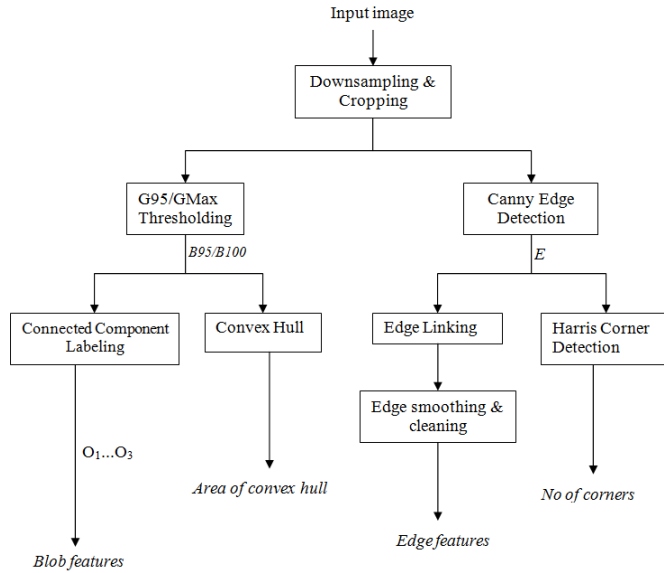


Fig. 2: Component diagram for image processing and feature extraction

experiments, the original size of the images is 2560x1920 pixels. We reduce the image size by 8-fold to get 320x240 sized image. Our analysis shows that the down-sampled images contain sufficient detail for feature extraction.

3.2 Image binarization

Image binarization is a technique for separating foreground and background regions in an image. For the protein images consisting of crystals, the crystal regions are expected to be represented as the foreground in the binary image. Images vary depending on crystallization techniques and imaging devices. This makes it difficult to use a fixed threshold for binarization. Therefore, dynamic thresholding methods are preferred. Different thresholding techniques provide good results for different images. Hence, extracting features from multiple thresholding techniques can be helpful. We apply two percentile based thresholding methods. The implementation and results for each of these techniques are described next.

- 1) *95th Percentile of Green (G95)* - When green light is used as the excitation source for fluorescence based acquisition, the intensity of the green pixel component is observed to be higher than the red and blue components in the crystal regions [7]. We utilize this feature for image binarization. First, threshold intensity τ_{g95} is computed as the 95th percentile intensity of the green component in all pixels. This means that the number pixels in the image with the green component intensity below this intensity constitute around 95% of the pixels. Also, a minimum gray level intensity condition ($\tau_{min} = 40$) is applied. All pixels with gray

level intensity greater than τ_{min} and having green pixel component greater than τ_{g95} constitute the foreground region while the remaining pixels constitute the background region.

- 2) *Max green threshold (GMax)* - This technique is similar to the 95th percentile green intensity threshold described earlier. In this method, maximum intensity of green component (τ_{gmax}) is used as the threshold intensity for green component. All pixels with gray level intensity greater than τ_{min} and having green pixel component greater than τ_{gmax} constitute the foreground region while the remaining pixels constitute the background region. The foreground (object) region in the binary image from this method is usually smaller than the foreground region from G95 threshold.

Fig. 3 shows some sample thresholded images using the two methods. From the original and binary images in Fig. 3, we can observe that a single technique may not yield good results for all images. For the images (i) and (ii), the binary images with G95 provide better representation of the crystal objects. However, for image (iii), the result from GMax threshold provides better representation of the crystals.

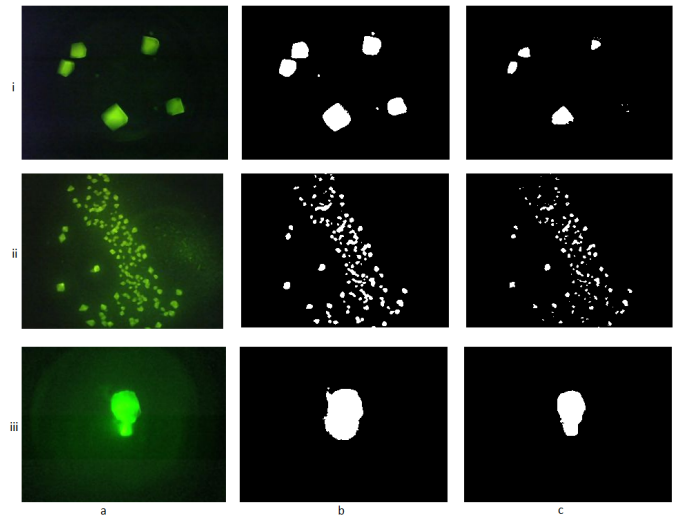


Fig. 3: Figure showing results of two image binarization techniques on crystallization trial images a) Original images b) G95 thresholded images c) GMax thresholded images

3.3 Image segmentation

After we generate the binary image, we apply connected component labeling to segment the regions (crystals). The binary image could be obtained from any of the thresholding methods. Let O be the set of the blobs in a binary image B , and B consists of n number of blobs. The blobs are ordered from the largest to the smallest such that $\text{area}(O_i)$

$\geq \text{area}(O_{i+1})$. Each blob O_i is enclosed by a minimum boundary rectangle (MBR) having width (w_i) and height (h_i). In our implementation, we define the minimum size of the blob to be 25 pixels.

We include the number of blobs in the binary image as one of the image feature. Likewise, for the 3 largest blobs (O_1 , O_2 and O_3), we extract the following features and append it to our feature vector.

- 1) *Blob area* - This is the area of the minimum bounding rectangle (MBR) enclosing the blob. In other words, it is simply the number of pixels in the blob image.
- 2) *Blob perimeter* - This is calculated as the sum of distance between each adjoining pair of pixels around the border of a blob.
- 3) *Blob filled area* - This is calculated as the number of white pixels in the blob.
- 4) *Blob eccentricity* - This measure corresponds to the ratio of the length of the MBR to the the width of the MBR. Eccentricity value lies between 0 and 1 where 0 is obtained when the blob is a circle and 1 is obtained when the blob corresponds to a line segment.

If a binary image contains less than 3 blobs, the value 0 is used for each of these features. It should be noted that the blobs may not necessarily represent crystals in an image. For such cases, the blob features may not be particularly useful for the classifier.

3.4 Convex hull area

In binary images, convex hull is the smallest set of points that forms a polygon shape, which contains the entire objects under consideration [14]. Convex hull points of an object indicates us the smallest number of enclosing object points which can be useful to detect boundaries of the object. We use area of convex hull as another image feature. This feature is useful to determine how the crystals are spread in the image.

3.5 Canny edge detection

Canny edge detection algorithm [15] is one of the most reliable algorithms for edge detection. The algorithm consists of four major steps. Firstly, Gaussian smoothing is done to reduce noise in the image. After Gaussian smoothing, intensity gradient of the image is calculated in different directions. Edge detection operators like Robers, Perwitt, Sobel are used to find the first derivative in the horizontal direction (G_y) and the vertical direction (G_x). Then edge gradient and direction are determined as follows:

$$g = \sqrt{G_x^2 + G_y^2} \quad (1)$$

$$\theta = G_y/G_x \quad (2)$$

After finding the edge gradient and direction, the edges which do not have local maximum are suppressed and classified as weak edges. Likewise, edges with local maximum

are classified as strong edges. If a weak edge is in the neighbor of a strong edge, then it is reclassified as strong edge. The strong edges and the reclassified weak edges form the complete edge image. The result of applying canny edge detector on three images is shown in Fig. 4. Our results show that for most cases, the shapes of crystals are kept intact in the resulting edge image.

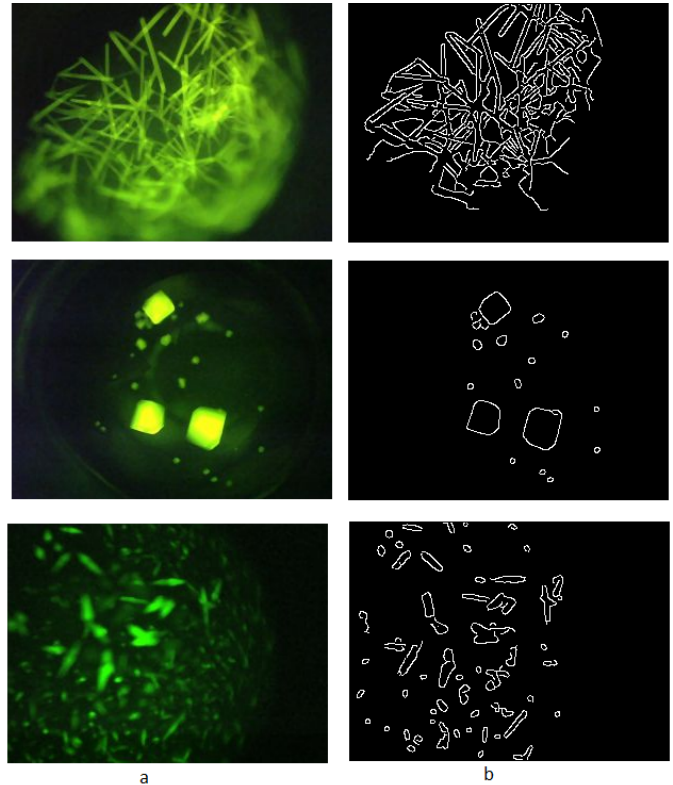


Fig. 4: Applying canny edge detection for 3 images a) Original image b) Canny edge image

3.6 Edge linking

An edge image can contain many edges which may or may not be part of the crystals. To analyze the shape and other edge related features, we link the edges to form graphs or segments. We used the MATLAB procedure by Kovese [16] to perform this operation. The input to this step is a binary edge image. Firstly, isolated pixels are removed from the input edge image. Next, the information of start and end points of the edges, endings and junctions are determined. From every end point, we track points along an edge until an end point or junction is encountered, and label the image pixels.

The result of edge linking is shown in Fig. 5c and Fig. 6c. The corresponding edge images are provided in Fig. 5b and Fig. 6b respectively.

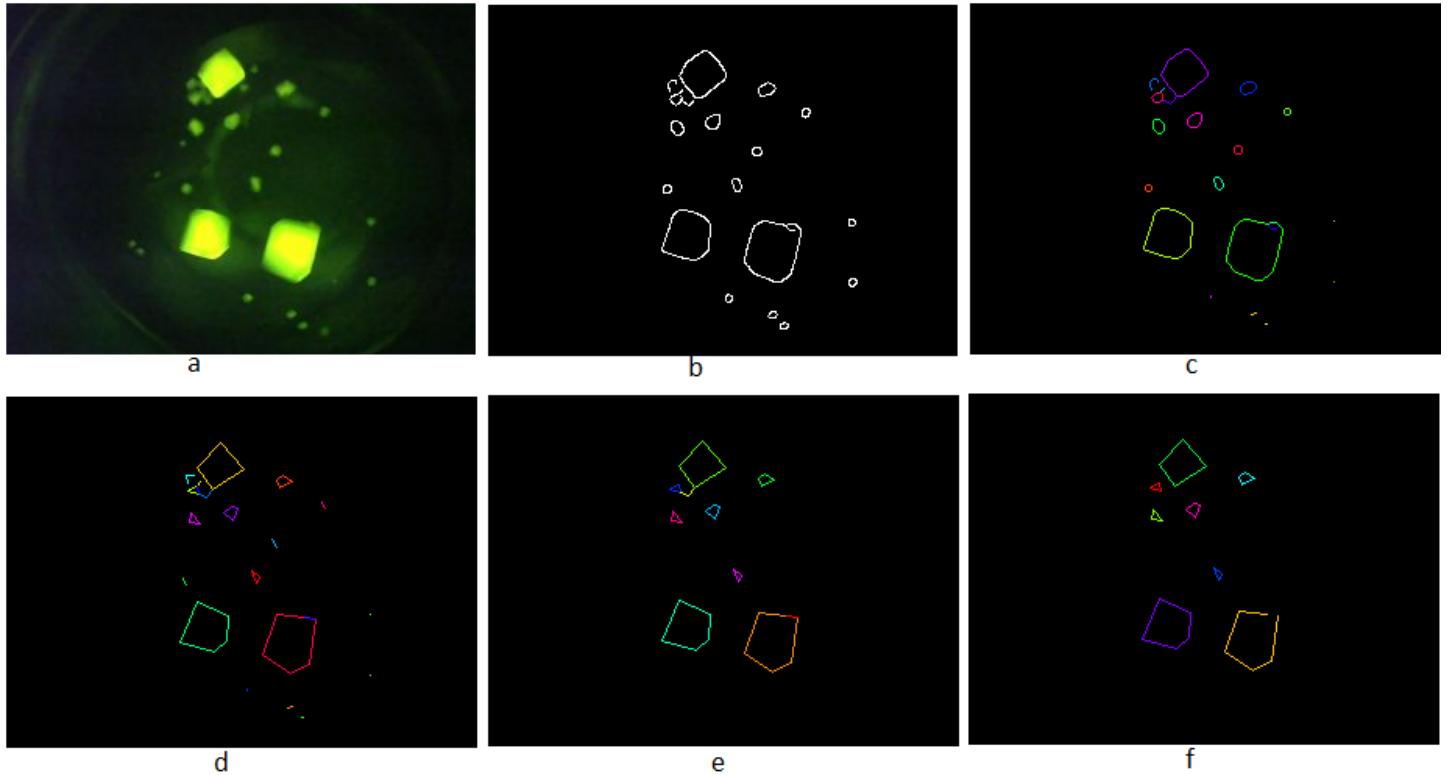


Fig. 5: Figure showing edge detection and edge feature extraction a) Original image b) Canny edge image c) Edge linking d) Line fitting e) Edge cleaning f) Image with cyclic graphs or edges forming line normals

3.7 Line fitting and edge cleaning

Due to problem with focusing, many edges could be formed. To reduce the number of edges and to link the edges together, line fitting is done. In this step, edges within certain deviation from a line are connected to form a single edge. The result from line fitting is shown in Fig. 5d and Fig. 6d. Here, the margin of 3 pixels is used as the maximum allowable deviation. From the figures, we can observe that after line fitting, the number of edges is reduced and the shapes resemble to that of exact shapes of the crystals. However, although desirable, this may not be achieved in all images.

Likewise, isolated edges and edges that are shorter than a minimum length are removed. The result from removing the uninterested edges is shown in Fig. 5e and Fig. 6e. Thus obtained list of edges is used to extract the following edge related features.

- 1) *Length of edges* - We determine the length of each edge using Euclidean distance measure. For an edge with the edge points (x_1, y_1) and (x_2, y_2) , the length (l) of the edge is computed using equation (3).

$$l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3)$$

- 2) *Angle between the edges* - We determine the slope of

each line and use it to compute the angle between connected lines. If two adjacent lines are almost perpendicular to each other, that provides a hint for the object to be small crystal or large crystal.

- 3) *Line normals* - Two lines are said to form line normals if the angle between the lines is 90 degrees. For each connected edge segment, we determine if two edges are perpendicular with each other. We consider two lines to be normals if the angle between the lines θ lies between 60 and 90 i.e., $60 \leq \theta \leq 120$.
- 4) *Cyclic graphs* - We check the edge link list and determine if the edges form a cycle. This is a useful feature to distinguish between needle crystals and other crystals.

Fig. 5f and Fig. 6f provide the edge linked image with only the edge segments that are cyclic or have line normals.

3.8 Harris corner detection

Corner points are considered as one of the uniquely recognizable features in an image. A corner is the intersection of two edges where the variation in both x and y gradient vector directions is very high. Harris corner detection [17] exploits this idea and it basically measures the change in intensity of a pixel (x, y) for a displacement of a search window in all directions. We apply Harris corner detection

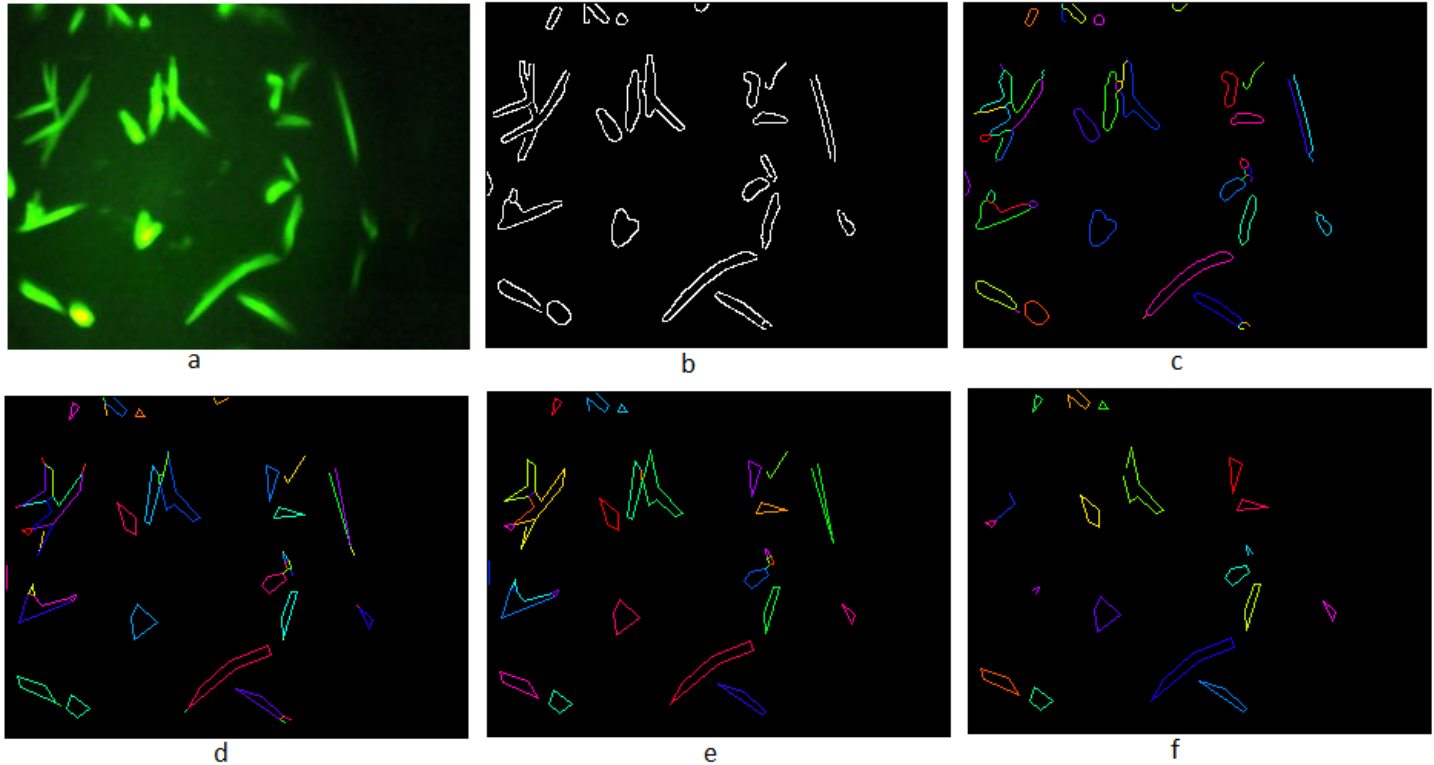


Fig. 6: Figure showing edge detection and edge feature extraction a) Original image b) Canny edge image c) Edge linking d) Line fitting e) Edge cleaning f) Image with cyclic graphs or edges forming line normals

and count the number of corners as the image feature.

3.9 List of features

For each image, we apply 2 dynamic image thresholding methods. Connected component labeling is done on the thresholded images and corresponding blob features are extracted. From each binary image, we extract $3 \times 4 + 2 = 14$ blob features. Likewise, we apply canny edge detection and extract 11 edge and corner features. Therefore, we extract a total of $2 \times 14 + 11 = 39$ features per image. Below is the list of all the extracted features.

- 1) Blob features
 - a) Area of the 3 largest blobs
 - b) Perimeter of the 3 largest blobs
 - c) Filled area of the 3 largest blobs
 - d) Eccentricity of the 3 largest blobs
 - e) No of blobs
 - f) Area of convex hull
- 2) Edge features
 - a) No of segments (graphs)
 - b) No of 1 edge graphs
 - c) No of 2 edge graphs
 - d) Has cyclic graph (0 or 1)
 - e) Has line normals (0 or 1)
 - f) No of cyclic graphs

- g) No of graphs with line normals
- h) Average length of edge in all segments
- i) Sum of lengths of all edges
- j) Maximum length of an edge
- k) No of Harris corner points

4. Experimental Results

Our experimental dataset consists of 212 expert labeled images. The images are hand-labeled by an expert into 4 different categories - Needle Crystals (NC), Small Crystals (SC), Large Crystals (LC) and Other Crystals (OC). These are represented in the proportion 24%, 20%, 35% and 21% respectively. Each image is processed as described in the earlier section and 39-dimension feature vector is obtained by extracting the blob, edge and corner features. We use decision tree as the classifier and evaluate the performance using 10-fold cross validation. Table 1 provides the resulting confusion matrix. We are able to achieve an accuracy of 75% $[(38+36+58+26)/212]$ on average for a four-class classification problem.

Among the 4 classes, we can observe that the system distinguishes the small crystals and needle crystals with high accuracy. Distinction between large crystals and other crystals is the most problematic.

From our discussion with the expert, small and large crystals are the most important crystals in terms of their

Table 1: Confusion Matrix

Actual Class	Observed Class			
	OC	NC	SC	LC
OC	26	4	4	10
NC	6	38	5	2
SC	1	3	36	3
LC	10	2	4	58

usability for the diffraction process. Therefore, it is critical not to misclassify the images in these categories into the other two categories. From Table 1, we can observe that our system misses 4 Small Crystals (1 image grouped as other crystals and 3 images grouped as needles). Likewise, our system classifies 10 Large crystals as Other Crystals and 2 Large Crystals as Needles. In overall, our system misses 16 critical images. Thus, the rate of miss of critical crystals of our system is around 8% [16/212]. This is a promising achievement for crystal subclassification of crystal categories.

5. Conclusion and Future Work

In this paper, we described a method for classifying different types of protein crystals in protein crystallization trial images. We extracted features related to edge and the shape characteristics of high intensity regions (blobs). We applied decision tree to develop the classification model and tested our experiments using 10-fold cross-validation. Our results indicate that the proposed classification technique produces a reasonable classification performance.

Crystallographers can not fully rely on the system as the classification accuracy is not very high. Hence, we need to improve the accuracy. The performance of our system depends on the accuracy of image binarization. In some images, the thresholded images do not capture the shapes of crystals correctly. Therefore, the features extracted from blobs may not necessarily represent crystals. Because of this, the features extracted from those blobs are not useful. To solve this problem, we plan to investigate different thresholding techniques. Our initial study shows that using the best thresholded image for feature extraction improves the classification performance.

We also plan to investigate hierarchical classification to obtain the decision model for the classification problem.

6. Acknowledgement

This research was supported by National Institutes of Health (GM090453) grant.

References

- [1] M. L. Pusey, Z.-J. Liu, W. Tempel, J. Prassman, D. Lin, B.-C. Wang, J. A. Gavira, and J. D. Ng, "Life in the fast lane for protein crystallization and x-ray crystallography," *Progress in Biophysics and Molecular Biology*, vol. 88, no. 3, pp. 359 – 386, 2005.
- [2] C. A. Cumbaa, A. Lauricella, N. Fehrman, C. Veatch, R. Collins, J. Luft, G. DeTitta, and I. Jurisica, "Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates," *Acta Crystallographica Section D: Biological Crystallography*, vol. 59, no. 9, pp. 1619–1627, 2003.
- [3] C. Cumbaa and I. Jurisica, "Automatic classification and pattern discovery in high-throughput protein crystallization trials," *Journal of structural and functional genomics*, vol. 6, no. 2-3, pp. 195–202, 2005.
- [4] I. M. Berry, O. Dym, R. Esnouf, K. Harlos, R. Meged, A. Perrakis, J. Sussman, T. Walter, J. Wilson, and A. Messerschmidt, "Spine high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects," *Acta Crystallographica Section D: Biological Crystallography*, vol. 62, no. 10, pp. 1137–1149, 2006.
- [5] S. Pan, G. Shavit, M. Penas-Centeno, D.-H. Xu, L. Shapiro, R. Ladner, E. Riskin, W. Hol, and D. Meldrum, "Automated classification of protein crystallization images using support vector machines with scale-invariant texture and gabor features," *Acta Crystallographica Section D: Biological Crystallography*, vol. 62, no. 3, pp. 271–279, 2006.
- [6] M. J. Po and A. F. Laine, "Leveraging genetic algorithm and neural network in automated protein crystal recognition," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 1926–1929.
- [7] M. Sigdel, M. L. Pusey, and R. S. Aygun, "Real-time protein crystallization image acquisition and classification system," *Crystal Growth Design*, vol. 13, no. 7, pp. 2728–2736, 2013.
- [8] K. Saitoh, K. Kawabata, and H. Asama, "Design of classifier to automate the evaluation of protein crystallization states," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 1800–1805.
- [9] G. Spraggon, S. A. Lesley, A. Kreusch, and J. P. Priestle, "Computational analysis of crystallization trials," *Acta Crystallographica Section D: Biological Crystallography*, vol. 58, no. 11, pp. 1915–1923, 2002.
- [10] C. A. Cumbaa and I. Jurisica, "Protein crystallization analysis on the world community grid," *J Struct Funct Genomics*, vol. 11, no. 1, pp. 61–9.
- [11] J. Hung, J. Collins, M. Weldetsion, O. Newland, E. Chiang, S. Guerrero, and K. Okada, "Protein crystallization image classification with elastic net," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2014.
- [12] K. Saitoh, K. Kawabata, S. Kunimitsu, H. Asama, and T. Mishima, "Evaluation of protein crystallization states based on texture information," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3. IEEE, 2004, pp. 2725–2730.
- [13] X. Zhu, S. Sun, and M. Bern, "Classification of protein crystallization imagery," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1. IEEE, 2004, pp. 1628–1631.
- [14] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software (TOMS)*, vol. 22, no. 4, pp. 469–483, 1996.
- [15] J. Canny, "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 679–698, 1986.
- [16] P. D. Kovesei, "MATLAB and Octave functions for computer vision and image processing," CETSE Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia, available from: <<http://www.csse.uwa.edu.au/~pk/research/matlabfns/>>.
- [17] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

A Content Based Image Retrieval Approach Based On Document Queries

M. Ilie¹

¹Department Name, "Dunarea de Jos" University of Galati, Faculty of Automatic Control, Computers, Electrical and Electronics Engineering, Galati, Romania

Abstract - This paper presents a new content based image retrieval (CBIR) approach, which makes use of descriptors originating in the local and global search spaces. The algorithms extracts four colour descriptors, one texture descriptor and two local descriptors which are used to train the corresponding classifiers, based on neural networks. Subsequently, the classifiers are grouped in two weighted majority voting modules, for local and global characteristics. The system is tested on regular images and on document scans obtained from two datasets used a benchmarks in previous conferences, in order to verify the architecture robustness. The experimental results demonstrate the effectiveness of the proposed model.

Keywords: CBIR, neural networks, image descriptors, weighted majority voting

1 Introduction

The necessity of the content based image retrieval phenomenon was imposed by the problems encountered in different areas. Initially, the image classification was done based on text labels, which was proven to be very time consuming and error prone. Starting from this problem, the image processing techniques have been improved, combined and extended across a vast number of fields, like duplicate detection and copyright, creating image collections, medical applications, video surveillance and security, document analysis, face and print recognition, industrial, military and so on. The term of "content" implies that the images are deconstructed into descriptors, which are analyzed and interpreted as image features, as opposed to image metadata, like annotations, geo-tags, file name or camera properties (flash light on/off, exposure etc.)

The traditional CBIR approaches try to solve this problem by extracting a set of characteristics from one image and comparing it with another one, representing a different image. The results obtained until now are promising but still far from covering all the requirements risen by a real world scenario. Also, the current approaches target specific problems in the image processing context. Because of that,

most of the CBIR implementations work in a rather similar way, on homogenous data. This causes significant performance drops whenever the test data originates from a different area than the training set.

This paper proposes a CBIR architecture model with descriptors originating in different search areas. In order to be able to classify images originating in document scans, we have added an extra module, responsible for the document image segmentation stage. The user is offered the possibility of querying the engine with both document scans and regular images in order to retrieve the best N matches.

During the implementation stage we have faced multiple problems, as specified below:

- image preprocessing;
- extraction of characteristics from various spaces; implementation of a supervised machine learning module;
- document image segmentation;
- benchmarking the overall performance.

We have reached the conclusion that a CBIR engine can obtain better results in the presence of multiple sets of descriptors, from different search spaces or from the same one, even if the test images originate in very different areas.

2 Related work

The CBIR engines are trying to mimic the human behaviour when executing a classification process. This task is very difficult to accomplish due to a large series of factors.

The CBIR queries may take place at different levels [1]:

- feature level (find images with X% red and Y% blue);
- semantic level (find images with churches);
- affective level (find images where a certain mood prevails). There is no complete solution for the affective queries.

All the CBIR implementations use a vector of (global or local) characteristics which originate in different search spaces - colour, texture or shape.

In the colour space there are many models but recently the focus is set on various normalizations of the RGB one in the attempt of obtaining invariant elements. Two of the most interesting ones are c1c2c3 (which eliminates both shadows and highlight areas) and l1l2l3 (which eliminates only the shadows, but keeps the highlight areas) [2].

There are 4 large categories for determining the texture descriptors [3]:

- statistical (Tamura features, Haralick's co-occurrence matrices);
- geometrical (Voronoi tessellation features);
- spectral (wavelets [4], Gabor and ICA filters);
- model based (MRFs [5], fractals).

One of the most widely embraced approach is to use local binary patterns [6].

In what regards the local descriptors, probably the most famous algorithm (scale invariant feature transform SIFT) was introduced by David Lowe [7]. Since then, many approaches have been developed. Some of the most popular ones are based on speeded un robust feature (SURF) [8], histogram of oriented gradients HOG [9], gradient location and orientation histogram (GLOH) [10] or local energy based shape histogram (LESH) [11].

3 Our approach

The proposed approach targets to classify a mixed set of images, containing real world scenes and document scans. The system mainly follows the standard CBIR architecture as it can be seen in the image below. It is composed of two interconnected submodules:

- the training and learning module;
- the document classification module.

A valid use case scenario contains the below stages:

- the system is trained on a set of images;
- each image is analyzed and decomposed in relevant descriptors;
- the descriptors are provided as input to a machine learning module, which is in charge of setting the class boundaries;
- each new regular image (not document) is decomposed and classified accordingly;
- each new document scan is preprocessed and segmented. The extracted images are then classified;
- the system extracts the 10 most relevant results and provides them as an answer to the user query.

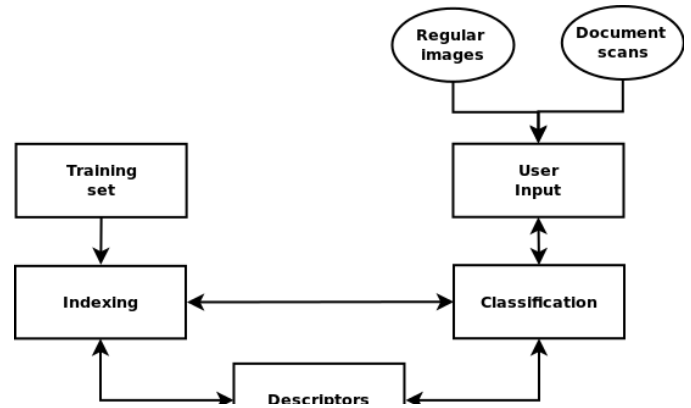


Figure 1. The basic system architecture

The indexing process is based on supervised machine learning and is conducted on regular images. The user is allowed to enter queries based on both image types.

We are using a mixed set of image characteristics:

- different colour spaces;
- texture space;
- local descriptors.

We have not used any shape descriptors, as the preliminary tests showed that in this area these do not produce a noticeable improvement. The main problem was caused by the fact that the objects contained in the images may be affected by problems like occlusion or clutter.

In the colour descriptors area, we have used 4 sets of characteristics, as it follows:

- c1c2c3 and l1l2l3. As explained above, these colour spaces are very useful when applied on real world images. The coordinates are described by the equations below:

$$c_1 = \arctg \frac{R}{\max(G,B)}; \quad (1)$$

$$c_2 = \arctg \frac{G}{\max(R,B)}; \quad (2)$$

$$c_3 = \arctg \frac{B}{\max(G,R)}; \quad (3)$$

$$l1(R, G, B) = \frac{(R - G)^2}{(R - G)^2 + (R - B)^2 + (B - G)^2} \quad (4)$$

$$l2(R, G, B) = \frac{(R - B)^2}{(R - G)^2 + (R - B)^2 + (B - G)^2} \quad (5)$$

$$l3(R, G, B) = \frac{(B - G)^2}{(R - G)^2 + (R - B)^2 + (B - G)^2} \quad (6)$$

- the whole image in RGB coordinates;
- the RGB histogram, with 256 bins.

Each of the four sets of characteristics is used as an input for a standard feed forward/back propagation neural network. The neural networks' outputs are then collected by a simplified weighted majority voting module, as it can be observed in the image below.

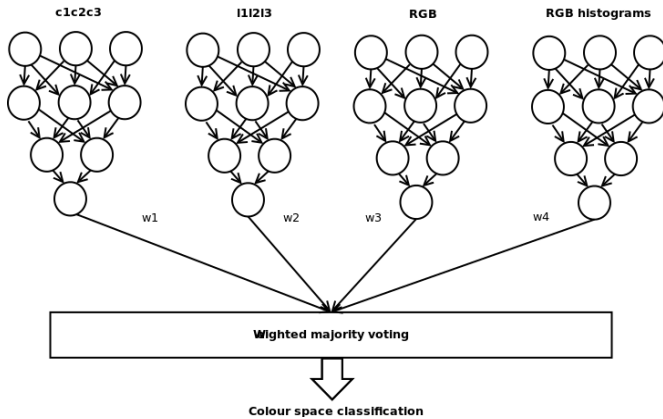


Figure 2. Colour space classification

The weighted module works according to the below algorithm:

- let n be the number of accepted classes and k be the number of classifiers;
- each neural network will produce on the final layer a vector $C_x = \{c_1, c_2, \dots, c_n\}$, where $1 \leq x \leq k$;
- the weight associated to the output layer will be $W = \{w_1, w_1 \dots, w_k\}$;
- the weighted result will be provided by the $w_i C_i$ sum, as specified below, where $R \in [1, n]$, $\max(C)$ represents the maximum value obtained for a certain class, and idx represents the position of this class in the final vector

$$R = idx \left(\max \left(\sum_{i=1}^k w_i C_i \right) \right) \quad (7)$$

In the texture space area we have chosen an approach based on local binary pattern descriptors, mainly because of their invariant properties for colour or rotation.

For the local descriptors we have chosen two sets of characteristics, based on scale invariant feature transform (SIFT) and histogram of oriented gradients (HOG). Traditionally, the HOG descriptors are used in order to train an SVM classifier, but since we are dealing with a multiple classification problem, we have used neural networks in the learning stage for both descriptors. The two types of local descriptors produced similar results during the tests, therefore the combined classifier for SIFT and HOG uses equal weights of 50%.

The final classifier includes an additional weighted majority voting module, as shown by the image below.

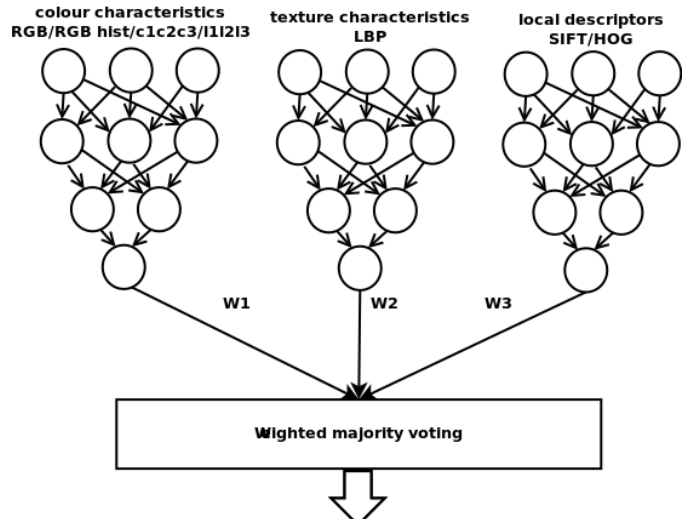


Figure 3. The final classification

Since the aim is to classify document scans as well as regular images, we have also included an additional document analysis module. Its purpose is to process the scans and extract the images included in the document, in order to pass them subsequently to the module in charge of extraction of descriptors and to classify them accordingly. We are not interested in text segmentation, therefore this module will only binarize the document and go through a bottom-up [12] image segmentation stage, based on the below steps:

- text filtering, implemented as a simplified XY axis projection module [13];
- the document is split in tiles, which are analyzed according to their average intensity and variance. The decision criteria is that in a particular tile, an image tends to be more uniform than the text;
- the remaining tiles are clustered through a K-Means algorithm, which uses as a decision metric the Euclidean distance;
- the clusters are filtered according to their connectivity and scarcity scores in order to eliminate tiles containing text areas with different fonts, affected by noise/poor illumination or by page curvature;
- the final clusters are exposed to a reconstruction stage and merged into a single image, which is provided as an input to the modules in charge of descriptors extraction/classification.

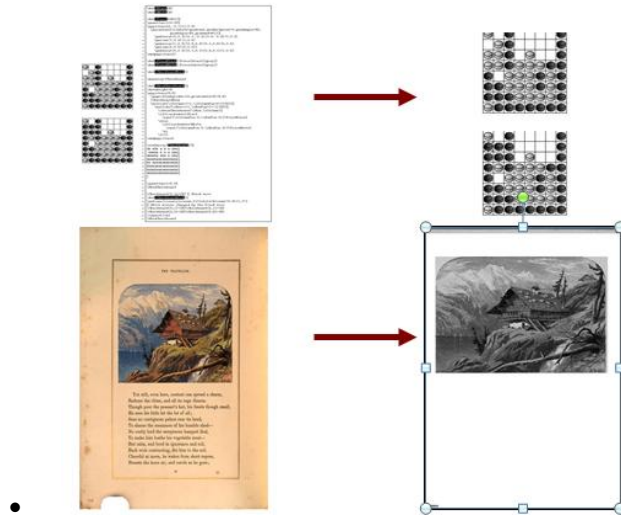


Figure 4: Document image segmentation results

All the neural networks have the same structure. The transfer function is sigmoid and the images in the training set have been split in three groups:

- 60% for training;
- 20% for cross-validation;
- 20% for testing.

In order to validate the neural network progress, we have used gradient checking on the cost function specified below:

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m \sum_{k=1}^K (y_k^{(i)} \log(h_{\theta}(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - h_{\theta}(x^{(i)}))_k) \right] + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\theta_{ji}^{(l)})^2 \quad (8)$$

with the following notations:

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

– the (input, output) vector

$$h_{\theta}(x) \in R^K \text{ – the hypothesis function}$$

L - the number of layers

sl - the number of neurons in a specific layer l

K - the number of classes, with $y \in R^K$

Θ - the matrix which stores the weights for each layer

4 Experimental setup

We aimed to create a scalable application, portable between different architectures and operating systems. Therefore, in what regards the programming language, we have chosen python over Matlab or Octave, for practical reasons, especially for the libraries which facilitate the user interface generation, socket management and data processing. The operating system is a 12.04 LTS 32 bit Ubuntu, running on a machine with two cores with hyper threading and 4 GB of RAM. The software architecture is modular to facilitate any subsequent refactoring; each sub-module is implemented in a class. In order to make better use of the hardware, the modules that require resources intensively use multi-threading and multi-processing techniques. The data is stored

in a MySQL database, based on MyISAM. The application follows the standard client-server architecture, in order to facilitate the exposure of functionalities to multiple users at once. So far, we have disregarded the user management problems.

We have restricted the number of recognized classes at 10 so far. The training was conducted on a CIFAR data set, provided by [14]; it includes 60000 small (32x32 pixels) colour images. The author's tests involved feed forward neural networks as well, with performances revolving around 87%.

The document scans data consisted of 1380 images, obtained from 2 sources:

- scans of old, degraded documents, used as a benchmark in the ICDAR 2007 conference [15];
- high quality copies, containing mostly manuals and documentation for the Ubuntu 12.04 operating system. In order to be able to use them, we have previously converted them from the pdf format to the jpeg one.

Initially we have tried replicating the CIFAR benchmark results. We have also used a neural network approach, based on RGB descriptors only; we obtained similar performances (85%). However, when we tried to classify an image originating from a different image set (document scans), the accuracy dropped significantly, by more than 10%. Therefore, we started experimenting with various combinations of RGB/c1c2c3/111213 descriptors. The results are described by the table below:

Table 1. Colour space experiments

Combined descriptors	Results
RGB+111213	82%
RGB+c1c2c3	84%
RGB+RGB histograms	69%
RGB	71%

As we can observe, the presence of the c1c2c3 and 111213 colour spaces produces an improvement of over 10%, leading to the below conclusions:

- the experiments conducted on real world images confirm the necessity of additional colour space descriptors;
- c1c2c3 produces the most solid performance boost. After analyzing the images in the data set, we have observed the presence of many pictures with shady areas, which shows that this colour space is adequate for these conditions. The image below shows the effects of the c1c2c3 normalization on an image containing highlight and shadow areas;
- introducing the RGB histograms as a global descriptor actually produced a performance drop, as two different images may have very similar colour histograms, yet a very different content. This was

mainly caused by the surrounding conditions in which the picture of a certain object was taken. Also, the presence of the shadow areas affect the histograms and implicitly the classification result.

After experimenting with various colour space weight values we have chosen the below values:

- the RGB histograms weights have been set to $w_4=10\%$, which improved the overall performance. We have kept this set of descriptors for situations where the colour plays a more important role in the classification process. In this case, the user will be able to adjust this value accordingly;
- the rest of the weights have been set to $w[1:3]=30\%$ (for the RGB/c1c2c3/111213 descriptors). This lead to a combined overall performance of 86%. As we mentioned before, the UI offers the user the possibility of manually adjusting the global colour relevance (associated to the RGB histograms) in the final result. As an example, the user can choose a combination like $w_1=20\%$, $w_2=20\%$, $w_3=20\%$ and $w_4=40\%$.

The next set of experiments was conducted in the texture space, with the LBP descriptor. The main problems in this area were related to choosing the cell shape and size, along with the number of pixels which compose the final descriptor. After a series of tests we concluded the below:

- choosing radial cells over square cells produces an overall performance increase of over 10%, going over 95%;
- the execution times are larger when using radial cells, especially due to the trigonometric calculus;
- over-increasing the cell size leads to performance drops, as the small textures are ignored.

The conclusion was that we will use square cells of 3x3 pixels and 8 pixels to compose the local texture descriptor.

Subsequently, we have started experimenting with the rest of the descriptors. For the HOG and SIFT algorithms we have used the authors' implementations. The tests involved the usage of singular descriptors and combining all of them together; the final weights have been set as it follows:

- $W_1=30\%$ for the colour space;
- $W_2=30\%$ for the texture space;
- $W_3=40\%$ for the local descriptors. These have been considered more representative than the global descriptors.

The results are presented in the table below:

Table 2. Experiments involving all descriptor spaces

Descriptor type	Results
Colour space	86%
Texture space (LBP)	85%
SIFT	82%

HOG	85%
All of the above	92%

The results show that combining multiple types of descriptors from multiple search spaces leads to performance improvements. On the above mentioned data set, the results are promising and show an increase of over 5%. Also, the proposed architecture is able to correctly classify images obtained from document scans as well as regular images.

5 Future research

In the future we would like to continue the research conducted so far. There are many areas which can be improved and also, upon refactoring they can provide new functionalities:

- we intend to add a module in charge of collecting an user score, which can be used later for altering the default weights in the majority voting module. This way, the CBIR engine will be able to provide more representative result for the user. Also, this feature can be combined with a user management module so that the system can recall the user's preferences;
- we also intend to insert a module that can analyze a certain image and compute how many shadow areas it contains. This module would help in deciding which colour space is more adequate for each particular situation;
- the number of characteristics is very large; only the SIFT descriptors may go over 100000 for 640x480 images. In order to be able to compute the results much faster, we considering the possibility of adding a module in charge of reducing the dimensionality;
- in the document processing area, the system is currently cropping out the images from a scan. In order to have more accurate results, we intend to add an OCR module, which can extract the text content as well. The text can be later on reduced to keywords, which can be used in the classification and retrieval process.

6 Acknowledgements

The authors would like to thank the Project SOP HRD /107/1.5/S/76822 - TOP ACADEMIC, of University "Dunarea de Jos" of Galati, Romania.

7 References

- [1] Sebe, N. Feature extraction & content description - DELOS - MUSCLE Summer School on Multimedia digital libraries, Machine learning and cross-modal technologies for access and retrieval. www.videolectures.net. [Online] 02 25, 2007. www.videolectures.net/dmss06_sebe_fecfd/.
- [2] Colour-based object recognition. Gevers, T., Smeulders, A.W. s.l. : Pattern Recognition, 1999, Vol. 32.

- [3] Vassilieva, Natalia. RuSSIR - Russian Summer School in Information Retrieval . [Online] 2012. http://videlectures.net/russir08_vassilieva_cbir/.
- [4] Illumination invariant extraction for face recognition using neighboring wavelet coefficients. X. Cao, W. Shen, L.G. Yu, Y.L. Wang, J.Y. Yang, Z.W. Zhang. 2012, Pattern Recognition.
- [5] Range map superresolution-inpainting, and reconstruction from sparse data. Arnav V. Bhavsar, Ambasadram N. Rajagopalan. 2012, Computer Vision and Image Understanding.
- [6] Wikipedia. Local binary patterns. www.wikipedia.org. [Online] 11 08, 2011. http://en.wikipedia.org/wiki/Local_binary_patterns.
- [7] Object recognition from local scale-invariant features. Lowe, David. s.l. : International Conference on Computer Vision, 1999.
- [8] Herbert Bay, Tinne Tuytelaars, Luc Van Gool. Speeded-Up Robust Features (SURF). Zurich, Leuven, Belgia : s.n., 2008.
- [9] Histograms of Oriented Gradients for Human Detection. Navneet Dalal, Bill Triggs. 2005, International Conference on Computer Vision & Pattern Recognition, pp. 886-893.
- [10] Krystian Mikolajczyk, Cordelia Schmid. A Performance Evaluation of Local Descriptors. IEEE transactions on pattern analysis and machine intelligence. 2005.
- [11] Head Pose Estimation in Face Recognition across Pose Scenarios. Saquib Sarfraz, Olaf Hellwich. Madeira, Portugal : s.n., 2008. Proceedings of VISAPP 2008, Int. conference on Computer Vision Theory and Applications, pp. 235-242.
- [12] A survey of document image classification: problem statement, classifier architecture and performance evaluation. Nawei Chen, Dorothea Blostein. 2007, International Journal of Document Analysis and Recognition (IJ DAR), p. Volume 10.
- [13] Recursive X-Y Cut using Bounding Boxes of Connected Components. Jaekyu Ha, Robert M. Haralick.
- [14] Krizhevsky, Alex. Learning Multiple Layers of Features from Tiny Images. 2009.
- [15] A. Antonacopoulos, D. Bridson, C. Papadopoulos. ICDAR 2007 Page Segmentation Competition. ICDAR. [Online] 2007. http://www.primaresearch.org/ICDAR2007_competition/.

Enhanced Spatial Fuzzy C-Means Algorithm for Medical Image Segmentation

Myeongsu Kang¹, Jaeyoung Kim¹, Cheol-Hong Kim², and Jong-Myon Kim^{1,*}

¹Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, South Korea

²School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea
ilmareboy@ulsan.ac.kr, kim7097@mail.ulsan.ac.kr, chkim22@chonnam.ac.kr, jmkim07@ulsan.ac.kr

Abstract - Image segmentation is an essential process in image analysis and is mainly used for automatic object recognition. Fuzzy c-means (FCM) is one of the most common methodologies used in clustering analysis for image segmentation. FCM clustering measures the common Euclidean distance between image pixels based on the assumption that each pixel has equal importance. However, image pixels are highly correlated, and thus it is necessary to exploit the spatial information to further improve clustering quality and correct misclassified pixels from noisy regions. To deal with this issue, this paper proposes an enhanced spatial FCM (ESFCM) that takes into account the influence of neighboring pixels on the center pixel by assigning weight to the neighbors by utilizing both pixel intensities and locations. In addition, the proposed ESFCM is robust to impulsive noise due to calculating the membership function of FCM using the vector median in a spatial domain. Experimental results indicate that the proposed ESFCM outperforms other FCM clustering algorithms using spatial information such as spatial fuzzy c-means and spatial fuzzy c-means modified in terms of clustering quality. In addition, the proposed ESFCM is more robust to impulsive noise than the other FCM clustering algorithms.

Keywords: Enhanced spatial fuzzy c-means, fuzzy c-means, medical image segmentation, vector median

1 Introduction

Image segmentation is an important analysis step for computer-aided diagnosis and therapy [1, 2]. This process separates an image into distinct classes such as brain tumors, edema, and necrotic tissues, enabling early detection of abnormal changes in tissues and organs by quantifying tissue volumes. The field of medicine has become an attractive domain for the application of fuzzy set theory. Fuzzy sets were introduced in 1965 by Lotfi Zadeh to merge mathematical modeling with human knowledge in the engineering sciences [3]. Fuzzy models and algorithms for

pattern recognition are widely used in advanced information technology [4]. One of the most well-known methodologies in clustering analysis is fuzzy c-means (FCM) clustering which was proposed by Dunn *et al.* in 1974 and extended by Bezdek in 1981 [5].

The standard FCM utilizes the Euclidean distance between pixels for computing memberships in order to segment an image based on the assumption that each pixel has equal importance; this affects performance degradation of clustering in cases in which neighboring pixels have strong correlation such as magnetic resonance (MR) images [6]. Likewise, the conventional FCM fails to segment images corrupted by noise despite the fact that it performs well on noise-free images. To address these drawbacks, many improved FCM clustering approaches that incorporate local spatial contextual information in images have been proposed. This is useful for reducing noise distortion and intensity inhomogeneity on image segmentation [7, 8]. The modified variant of FCM called spatial FCM clustering was proposed by Chuang *et al.*; it utilizes spatial information in the FCM membership function and is less sensitive to noise [9]. In spatial FCM, however, an equal weighting factor is given to the adjacent pixels in the predefined window, which results in inaccurate segmentation.

To address this problem, Chaudhry *et al.* utilized weight depending on the contribution of the pixel and that is determined by using the Euclidean distance between neighboring pixels in a predefined window [10]. This method reduced the effects of noise compared to spatial FCM clustering and resulted in the formation of more homogeneous clustering than that of spatial FCM. To further enhance the clustering performance compared to these modified variants of FCM, this paper proposes an enhanced spatial FCM (ESFCM) using weight based on both pixel intensities and pixel locations. Furthermore, this study utilizes the vector median for computing the membership function of the proposed ESFCM, which decreases the number of misclassified pixels due to the impulsive noise inherent in an image.

The rest of this paper is organized as follows. Section 2 briefly introduces the standard FCM, and Section 3 proposes a new variant of FCM. Section 4 validates the effectiveness of the proposed enhanced spatial FCM. Finally, Section 5 concludes this paper.

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. NRF-2013R1A2A2A05004566).

* Corresponding author.

2 Standard fuzzy c-means

FCM [5] is one of the most well-known methodologies in clustering analysis. Clustering is the process of portioning an image into regions (or classes) such that each region is homogeneous and none of the unions of two adjacent regions is homogeneous. FCM clustering is an iterative algorithm-based clustering technique that produces an optimal number of c partitions, with centroids $V = \{v_1, v_2, \dots, v_c\}$ which are exemplars, and radii which define these c partitions. Suppose the unlabeled dataset $X = \{x_1, x_2, \dots, x_n\}$ is the pixel intensity, where n is the number of image pixels whose memberships are to be determined. The FCM clustering process partitions the dataset X into c clusters. The objective function of the standard FCM is defined as follows:

$$J_m(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m d(x_k, v_i), \quad (1)$$

where $d(x_k, v_i)$ represents the Euclidean distance between pixel x_k and centroid v_i , and u_{ik} represents the fuzzy membership of the k th pixel with respect to cluster i with the constraint $\sum_{i=1}^c u_{ik}^m = 1$, and the degree of fuzzification $m \geq 1$.

The data point x_k belongs to a specific cluster i which is given by the membership value u_{ik} of the data point to that cluster. Local minimization of the objective function $J_m(U, V)$ is accomplished by repeatedly adjusting the values of u_{ik} and v_i according to the following equations:

$$u_{ik} = \left[\sum_{j=1}^c \left(\frac{d^2(x_k, v_i)}{d^2(x_k, v_j)} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (2)$$

$$v_i = \frac{\sum_{k=0}^n u_{ik}^m \cdot x_k}{\sum_{k=0}^n u_{ik}^m}, \quad 1 \leq i \leq c. \quad (3)$$

As $J_m(U, V)$ is iteratively minimized, v_i becomes more stable. The pixel clustering iterations are terminated when the termination measurement $\max_{1 \leq i \leq c} \left\{ \left\| v_i^{(t)} - v_i^{(t-1)} \right\| \right\} < \varepsilon$ is satisfied, where $v_i^{(t)}$ are the new centroids for $1 \leq i \leq c$, $v_i^{(t-1)}$ are the previous centroids for $1 \leq i \leq c$, and ε is a predefined termination threshold. The output of the FCM algorithm is the cluster centroids V and the fuzzy partition matrix $U_{C \times N}$.

3 Enhanced spatial fuzzy c-means

In spite of the fact that the image pixels are highly correlated and the spatial relationship of neighboring pixels is an important characteristic for image segmentation, the standard FCM clustering does not fully utilize this spatial information. To deal with this drawback, authors in [9, 10] used the spatial information of neighbors in a predefined window and increased the probability that a pixel in the predefined window belongs to the same cluster if its neighboring pixels belong to a certain cluster. Moreover, the spatial information is helpful for reducing the number of misclassified pixels due to noisy components in an image. To improve clustering performance compared to these modified FCM algorithms, this paper proposes an ESFCM that exploits more spatial information.

Unlike standard and modified FCM algorithms, this study finds the vector median of neighbors falling into a predefined window (3×3 window in this study) around x_k and utilizes the vector median for computing the membership of pixel x_k in order to correct misclassified pixels from noisy regions. Consequently, the membership, u_{ik} , in the proposed ESFCM is defined as follows:

$$u_{ik} = \left[\sum_{j=1}^c \left(\frac{d^2(VM(x_k), v_i)}{d^2(VM(x_k), v_j)} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad (4)$$

where $VM(x_k)$ is the vector median of the window around pixel x_k . To find the vector median, this study utilizes the cumulative distances criterion and determines the lowest-ranked input vector as the vector median. Let $N_k = \{x_1, x_2, \dots, x_n\}$ be a set of neighbors centered on the pixel x_k in the predefined window, where n is the dimension of the window. For every vector, x_i , in the window, the cumulative distances to all the other vectors using a norm metric are computed, resulting in

$$L_i = \sum_{j=1}^n \left\| x_i - x_j \right\|_2^2, \quad i = 1, 2, 3, \dots, n, \quad i \neq j. \quad (5)$$

The vector median, $VM(x_k)$, is associated with the input vector yielding the minimum accumulated distance. In general, MR images include impulsive noise, which is independent and uncorrelated with the image pixels. In addition, this noise is randomly distributed over an image. Thus, uncertainty is widely presented due to impulsive noise, which results in low clustering performance. The proposed membership function reduces the misclassified pixels due to the impulsive noise by exploiting the vector median instead

of the center pixel in the window. In addition, the ESFCM utilizes a neighbor-weighting coefficient, p_{ik} , to further improve the segmentation performance, which is defined as follows:

$$p_{ik} = \sum_{j=1}^n \left(\frac{u_{ij} \cdot d(L(x_k), L(x_j))}{d^2(x_k, x_j)} \right), \quad (6)$$

where $L(x_k)$ and $L(x_j)$ are the locations of pixels x_k and x_j in the window, respectively. Likewise, $d(L(x_k), L(x_j))$ is the Euclidean distance, which can be expressed by

$$d(L(x_k), L(x_j)) = \sqrt{(k_x - j_x)^2 + (k_y - j_y)^2}, \quad (7)$$

if $L(x_k) = (k_x, k_y)$, $L(x_j) = (j_x, j_y)$.

The neighbor-weighting coefficient using both pixel intensity and location information of a pixel for a cluster leads to a higher probability if the majority of its neighborhood belongs to the same clusters. In other words, the greater number of neighbors in the same cluster, the higher is the probability that the center pixel is in that cluster. The weighted coefficient function, $f(p_{ik})$, is incorporated into the membership function of the standard FCM, and a new membership function, w_{ik} , is defined as follows:

$$w_{ik} = \frac{u_{ik} \cdot f^{1/m-1}(p_{ik})}{\sum_{j=1}^c u_{jk} \cdot f^{1/m-1}(p_{jk})}. \quad (8)$$

Using the new membership function, w_{ik} , the centroid values, v_i , of the proposed ESFCM are computed such that

$$v_i = \frac{\sum_{k=1}^n w_{ik}^m \cdot x_k}{\sum_{k=1}^n w_{ik}^m}, \quad 1 \leq i \leq c. \quad (9)$$

4 Performance evaluation

4.1 Parameter setup

Initialization for the degree of fuzzification m is very important in FCM. FCM clustering produces terminal

partitions $\bar{U} = [1/c]$ when $m \rightarrow \infty$. In contrast, when $m \rightarrow 1$, this reduces to hard c-means and terminal partitions become more and more crisp. In the method of *Bezdek* [5], the authors experimentally determined the optimal interval for the degree of fuzzification and found it to range from 1.1 to 5. In this study, we selected the value of m as 2 so as to have an optimal balance of speed and accuracy for all of the FCM-based clustering algorithms. The termination threshold ε controls the duration of iteration as well as the optimal terminal partition of the fuzzy clustering. *Bezdek* [5] experimentally determined the optimal interval for the termination threshold and found it to range from 0.01 to 0.0001. In this study, we selected the termination threshold value to be 0.001.

The initialization of the centroid of a cluster is also important in FCM clustering because it is a searching technique that yields local maxima, thus greatly reducing the performance of clustering. In addition, when clustering is initialized from a different starting point, different solutions are found for the same terminal partition. In this study, the centroids were initialized by assigning the number of clusters (denoted as c), with points uniformly distributed according to the gray image (intensities ranging from 0 to 255).

4.2 Segmentation results

This paper evaluated the correctness of the segmentation using real brain scans with ground truth given by expert segmentations obtained from the Internet Brain Segmentation Repository (IBSR) website (<http://www.cma.mgh.harvard.edu/ibsr/>). A brain scan given is composed of a variable number of slices. We processed the slices individually. Figs. 1(a)–(b) show an example of the slices and the manual labeling provided by the IBSR site. A comparison of the segmentation results obtained by applying four clustering algorithms on a T1-weighted MR phantom is shown in Fig. 1. The segmentation results attained for two slices using four methods are shown in Figs. 1(c)–(f). In addition, the quantitative comparison scores corresponding to Fig. 1(a) for gray matter (GM) and white matter (WM) are given in Table 1, and comparison scores are computed as follows:

$$s_{i,j} = \frac{A_{i,j} \cap A_{ref,j}}{A_{i,j} \cup A_{ref,j}}, \quad (10)$$

where $A_{i,j}$ represents the set of pixels belonging to the j th class found by the i th algorithm, and $A_{ref,j}$ represents the set of pixels belonging to the j th class in the reference segmented image. From Fig. 1 and Table 1, it can be seen that the proposed ESFCM outperforms the conventional algorithms.

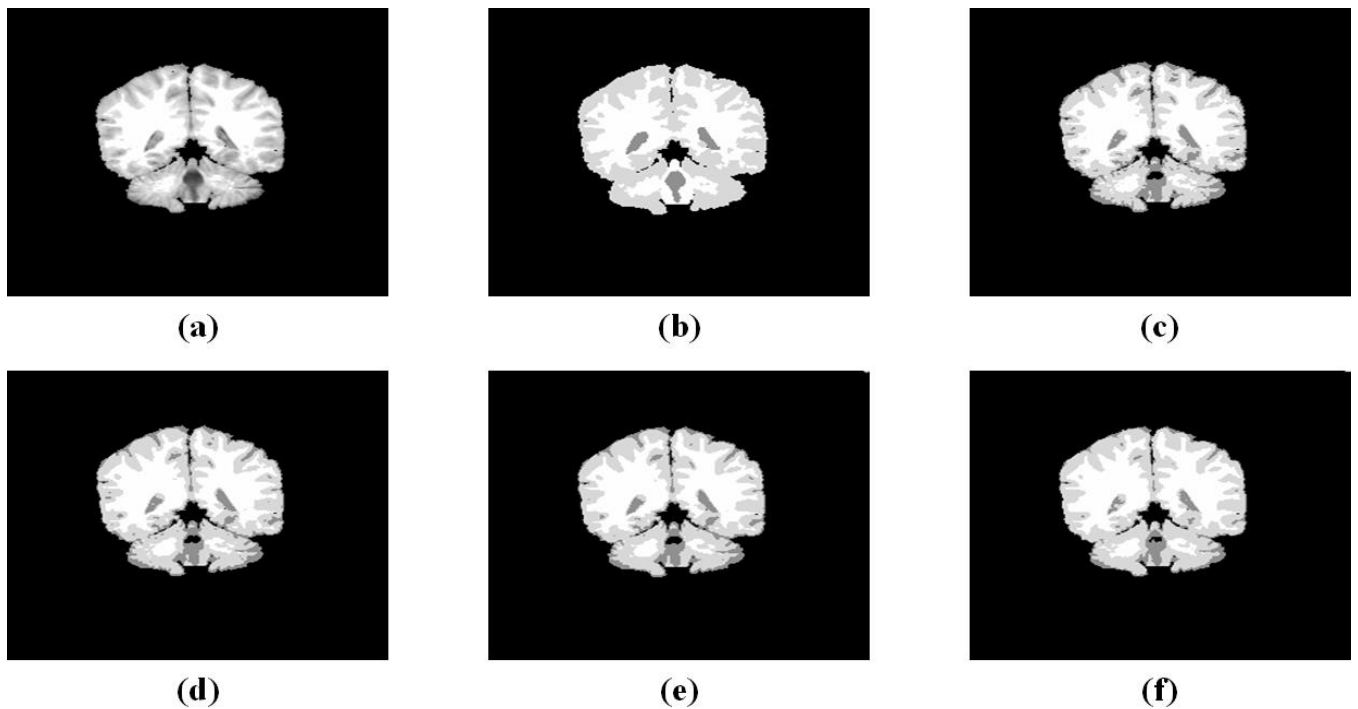


Fig. 1. Comparison of the segmentation results on a simulated brain MR image. (a) Original T1-weighted image, (b) manual class labeling of gray matter (GM) and white matter (WM) slice regions ; results obtained with (c) standard FCM, (d) spatial FCM [9], (e) spatial FCM modified [10], (f) the proposed ESFCM

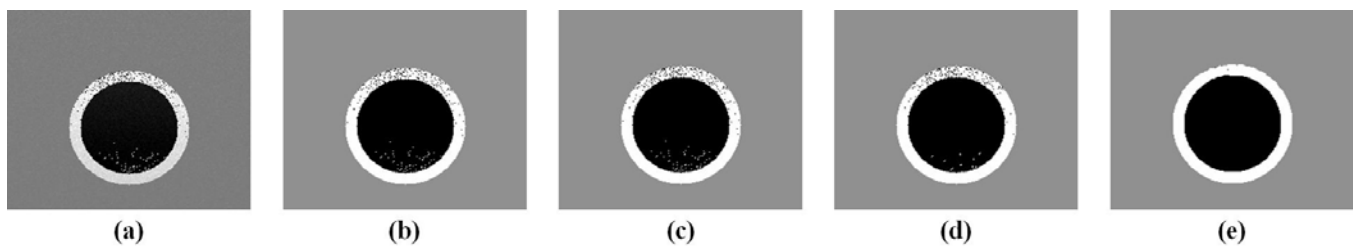


Fig. 2. (a) Synthetic circle image with small gradient noise, and resulting images of clustering using (b) standard FCM, (c) spatial FCM [9], (d) spatial FCM modified [10] and (e) the proposed ESFCM

Table 1 : Comparison scores of four segmentation approaches with Fig. 1(a)

Method	GM	WM
Standard FCM	0.6223	0.7870
Spatial FCM [9]	0.6478	0.7887
Spatial FCM modified [10]	0.6614	0.7833
Proposed ESFCM	0.7167	0.8014

We also simulated the effect of noise with target images for segmentation. In the conventional FCM, a noisy pixel can be wrongly classified due to its abnormal feature data. However, the proposed ESFCM can greatly reduce the effect of noise by incorporating spatial information. Fig. 2(a) depicts a synthetic circle image with small gradient noise

which is obtained from the IBSR website and Figs. 2(b)-(e) show the clustering results of four clustering methodologies. As shown in Fig. 2, the clustering result of the proposed ESFCM is superior to those of other algorithms.

5 Conclusions

FCM is one of the most well-known clustering algorithms. However, it uses the common Euclidean distance based on the assumption that each pixel has equal importance, resulting in clustering performance degradation since image pixels are highly correlated. To address this issue, this paper proposed ESFCM, which takes into account the influence of the neighboring pixels on the center pixel. In addition, this study utilized the vector median to reduce misclassified pixels due to the impulsive noise inherent in an image. Experimental results indicate that the proposed ESFCM significantly outperforms other FCM-based clustering algorithms.

6 References

- [1] S. Shen, W. Sandham, M. Grant, and A. Sterr, "MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 3, pp. 459–467, 2005.
- [2] K. S. Fu and J. K. Mu, "A survey on image segmentation," *Pattern Recognition*, vol. 13, pp. 3–16, 1981.
- [3] L. A. Zadeh, "Fuzzy sets," *Information Control*, vol. 8, pp. 338–353, 1965.
- [4] J. C. Bezdek, J. Keller, R. Krisnapuram, and N. R. Pal, "Fuzzy models and algorithms for pattern recognition and image processing." Springer, 1st edition, 2005.
- [5] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms." Plenum Press, New York, 1981.
- [6] S. R. Kannan, S. Ramathilagam, R. Devi, and E. Hines, "Strong fuzzy c-means in medical image data analysis," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2425–2438, 2012.
- [7] C. Qiu, J. Xiao, L. Yu, L. Han, and M. N. Iqbal, "A modified interval type-2 fuzzy c-means algorithm with application in MR image segmentation," *Pattern Recognition*, vol. 34, no. 12, pp. 1329–1338, 2013.
- [8] M. Gong, Y. Liang, J. Shi, W. Ma, and J. Ma, "Fuzzy c-means clustering with local information and kernel metric for image segmentation," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 573–584, 2013.
- [9] K. -S. Chuang, H. -L. Tzeng, S. Chen, J. Wu, and T. -J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics*, vol. 30, no. 1, pp. 9–15, 2006.
- [10] A. Chaudhry, M. Hassan, A. Khan, J. Y. Kim, and T. A. Tuan, "Image clustering using improved spatial fuzzy c-means," *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, Kuala Lumpur, Malaysia, February 2012, pp. 1–7.

Stereo Image Coding Scheme Based Upon Pattern Recognition

H. S. Li¹, Da Li², and Chen Li³

¹School of Information and Communication, Gulin University of Electronic Technology, Guangxi, China

²School of Information Science and Technology, Beijing Normal University, Beijing, China

³VMWARE, Inc, Palo Alto, California, USA

Abstract - To improve the coding performance of the left image, this paper presents a new stereo image coding scheme (SOM+PR+DE) based upon pattern recognition (PR) and disparity estimation (DE). We use pattern recognition (PR) to encode the left image instead of general JPEG algorithm. In our pattern recognition, Kohonen's self-organizing map (SOM) is used for the pattern library training. Disparity estimation is used to predict right image as usual. Both the PR and DE predicted error is coded by DCT and entropy coding. Experimental results on stereo image Pentagon show that the SOM+PR+DE algorithm can effectively improve the coding performance of the left image better than the JPEG+DE algorithm, when compression ratio is the same (6.5:1), the PSNR improvement is 2.78dB, and when PSNR is the same (30dB), compression ratio improvement is 1.9 times.

Keywords: Stereo image coding, self-organizing map, pattern recognition, disparity estimation

1 Introduction

Stereo image coding is one of the hotspots in the academe, schemes based upon wavelet and schemes combined with MPEG standard are the two main algorithms [1]-[7]. Most of the experts fix their attention on how to improve the veracity of disparity estimation and how to efficiently code the disparity estimation error. The scheme combined with MPEG standard is more popular, this predicts right image using disparity estimation (DE) technology and codes the left image and disparity estimation error using JPEG standard. The self-organizing feature maps (SOM) algorithm proposed by Kohonen is an efficient clustering algorithm, and has been extensively used in data compression and pattern recognition [8]-[13]. SOM algorithm is used to seek the optimal pattern library by training with a large number of sample sequences. we present a new coding scheme: the left image is coded by pattern recognition based upon SOM algorithm instead of the usual JPEG algorithm, the right image is predicted by disparity estimation, and the error of PR and DE are coded by DCT and Huffman coding. Experiments show that the SOM+PR+DE algorithm can effectively improve the coding performance of the left image compared with the JPEG+DE algorithm.

2 SOM+PR+DE Algorithm

2.1 Coding Scheme

Stereo Image coding scheme is described in Figure1. It can be summarized as follows

Step-0: Design an optimal pattern library using SOM algorithm by training a lot of stereo image samples.

Step-1: Divide the left image into blocks (for example 8×8), and search for its optimal matching pattern in the pattern library of each block, the matching pattern is called the predicted image block.

Step-2: The predicted error block obtained by subtracting the predicted image block from the original image block is coded by DCT, quantization and Huffman coding.

Step-3: Predict right image, using disparity estimation technology based on the reconstructed left image.

Step-4: Encode the disparity estimation error by DCT, quantization and Huffman coding.

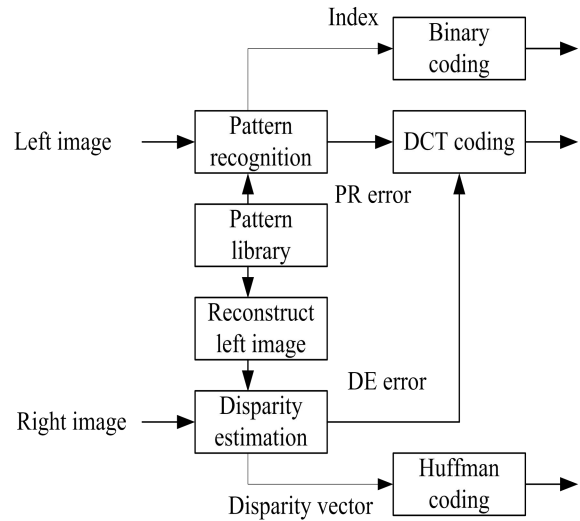


Figure 1 SOM+PR+DE coding scheme

2.2 SOM Algorithm

The SOM algorithm for designing a pattern library can be summarized as follows maximum allowed

Step-1: Given a neural network of size (N, M) , where N is the size of pattern library, and M is the size of each pattern. Choose a training vector set $\{\mathbf{X}_n, n = 0, 1, \dots, L-1\}$, initialize the pattern library $\{\mathbf{W}_j^{(0)}, j = 0, 1, \dots, N-1\}$ and the neighbourhood $\{NE_j^{(0)}, j = 0, 1, \dots, N-1\}$.

Step-2: Input training vector \mathbf{X}_n .

Step-3: Compute the distortion $d_j^{(n)}$ between input vector and each vector in the pattern library with some distortion measure, and select the winning vector j^* with minimum distortion.

Step-4: Modify the winning vector and its neighbouring vectors by

$$\mathbf{W}_j^{(n+1)} = \begin{cases} \mathbf{W}_j^{(n)} + \alpha(n)[\mathbf{X}_n - \mathbf{W}_j^{(n)}] & j \in j^*, NE_{j^*}^{(n)} \\ \mathbf{W}_j^{(n)} & else \end{cases} \quad (1)$$

$NE_{j^*}^{(n)} = A_0 + A_1 e^{-n/T_1}$, Euclidean distance neighbourhood

around the winning vector which is decreased with time.

Learning rate $\alpha(n) = A_2 e^{-n/T_2}$ is also decreased with time.

Step-5: let $n = n + 1$, go to Step-1.

2.3 Disparity Estimation and Coding

Disparity is the difference between the stereo image pairs. In the direction of the disparity vector, image pairs are highly similar. Disparity estimation which is used to predict the right image is similar with motion estimation. There are two main approaches of disparity estimation: block-based and mesh-based approach. Block-based disparity estimation can use fixed block and adaptive block, we used a fixed 8×8 block in this paper. In theory, stereo image pairs only have a disparity vector in the horizontal direction when the two cameras are parallel. Considering the difference between fact and theory, when search the matching block for right image blocks, we not only set large horizontal range, but also set a small vertical range. The PR error and DE error both are coded by DCT coding (including DCT, quantization, and Huffman coding).

2.4 Experimental Results

The stereo image pairs for codebook training and coding are both the standard testing sequence Pentagon, the resolution is $512 \times 512 \times 8$ bit, the size of the image block is 8×8 . Mean square error (MSE) is used as the distortion measure $d_j^{(n)} = \|\mathbf{X}_n - \mathbf{W}_j^{(n)}\|^2$. Image quality is measured by

PSNR, $PSNR = 10 \lg 255^2 / MSE$, where MSE is the mean square error between the original image and the coded image. The compression rate of the left image is $R_l = B_o / (B_c + B_e)$, where B_o is bits of original left image, B_c is bits of VQ address, B_e is bits of VQ error. The compression rate of the right image is $R_r = B_o / (B_v + B_e)$, where B_o is bits of original right image, B_v is bits of disparity vector, B_e is bits of disparity estimation error. Simple binary coding is used to code the VQ address, the size of the codebook is 1024, and thus one address needs 10 bits. We choose $\alpha(n) = e^{-n/8192}$ as the learning function and $NE_j^{(n)} = 0 + 4e^{-n/4096}$ as the neighbourhood function. The initial codebook is chosen from the training vector set randomly.

Compared with traditional JPEG+DE algorithm, SOM+PR+DE algorithm improves the performance mainly in left image coding. The comparison of the two algorithms is shown in Table 1. The experiment results show that SOM+PR+DE algorithm efficiently improves the coding performance of left image: when the compression rate is the same (about 6.5:1), PSNR improvement is 2.78dB, and when PSNR is the same (30dB), compression ratio improvement is 1.9 times.

Table 1 Comparison Results

JPEG+DE Algorithm		SOM+PR+DE Algorithm	
Compression rate	PSNR (dB)	Compression rate	PSNR (dB)
6.51:1	32.23	6.49:1	35.01
7.88:1	31.23	8.06:1	33.76
10.43:1	30.02	9.75:1	32.79
12.90:1	29.25	11.60:1	31.94
15.22:1	28.67	15.49:1	30.86
17.53:1	28.18	17.68:1	30.48
19.79:1	27.77	18.51:1	30.45
22.09:1	27.40	23.61:1	29.74
24.30:1	27.07	25.34:1	29.58
26.62:1	26.76	26.64:1	29.47

3 Conclusions

In this paper, we present a new stereo image coding scheme based on pattern recognition and disparity estimation technology. Experiment results show that this new scheme can improve coding performance of the left image efficiently. In the future work, we will make a study on how to get a more efficient pattern recognition library, improve the efficiency of error image coding, and make a Huffman coding table for disparity estimation error and pattern recognition error.

4 References

- [1] Yanwei Liu, Qingming Huang, Siwei Ma, Debin Zhao, Wen Gao. "Joint video/depth allocation for 3D video coding

- based on view synthesis distortion model”; *Signal Processing: Image Communication*, Vol.24, Issue 8, 666—681, Sep 2009.
- [2] M. Paul, Junbin Gao, M. Anotolovich. “3D motion estimation for 3D video coding”; *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1189—1192, 2012.
- [3] Wenyi Su, D. Rusanovskyy, M.M. Hannuksela, Houqiang Li. “Depth-based motion vector prediction in 3D video coding”; *IEEE International Conference on Picture Coding Symposium (PCS)*, 37—40, 2012.
- [4] Pei-Jun Lee, Xu-Xian Huang. “3D motion estimation algorithm in 3D video coding”; *International Conference on System Science and Engineering (ICSSE)*, 338—341, 2011.
- [5] C. Conti, J. Lino, P. Nunes, L. Ducla Soares, P. Lobato Correia. “Spatial prediction based on self-similarity compensation for 3D holoscopic image and video coding”; *IEEE International Conference on Image Processing*, 961—964, 2011.
- [6] C. Conti, J. Lino, P. Nunes, L.D. Soares. “Spatial and temporal prediction scheme for 3D holoscopic video coding based on H.264/AVC”; *International on Packet Video Workshop (PV)*, 143—148, 2002.
- [7] D.V.S.X. De Silva, W.A.C. Fernando, H.K. Arachchi. “A new mode selection technique for coding Depth maps of 3D video”; *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 686—689, 2010.
- [8] Aixia Yan, Xianglei Nie, Kai Wang, Maolin Wang. “Classification of Aurora kinase inhibitors by self-organizing map (SOM) and support vector machine (SVM)”; *European Journal of Medicinal Chemistry*, Available online 26, June 2012.
- [9] Wenxiu Zou, Asim Biswas, Xiaozeng Han, Bing Cheng Si. “Extracting soil water storage pattern using a self-organizing map”; *Geoderma*, Vol. 177–178, 18—26, May 2012.
- [10] H. Hikawa, K. Doumoto, S. Miyoshi, Y. Maeda. “Image compression with hardware self-organizing map”; *International Joint Conference on Neural Networks (IJCNN)*, 1—8, 2010.
- [11] I. Chaabouni, W. Fourati, M.S. Bouhlel. “An improved image compression approach with combined wavelet and self-organizing maps”; *IEEE Electro technical Conference (MELECON)*, 360—365, Online 4, Oct 2012.
- [12] E.J. Palomo, E. Dominguez. “Image compression based on growing hierarchical Self-Organizing Maps”; *International Joint Conference on Neural Networks (IJCNN)*, 1624—1628, 2012.
- [13] Teuvo Kohonen. “Essentials of the self-organizing map”; *Neural Networks*, Available online 4, Oct 2012.

SESSION

IMAGE ENHANCEMENT METHODS, NOISE REDUCTION ALGORITHMS, IMAGE QUALITY ASSESSMENT AND RELATED TECHNOLOGIES

Chair(s)

TBA

DENOISING CAMERA DATA: SHAPE-ADAPTIVE NOISE REDUCTION FOR COLOR FILTER ARRAY IMAGE DATA

Tamara Seybold[†] Bernd Klässner[‡] Walter Stechele[‡]

[†] Arnold & Richter Cine Technik, Türkenstraße 89, 80799 München, Germany

[‡] Technische Universität München, Arcisstraße 21, 80333 München, Germany

ABSTRACT

While denoising readily processed images has been studied extensively, the reduction of camera noise in the camera raw data is still a challenging problem. Camera noise is signal-dependent and the raw data is a color filter array (CFA) image, which means the neighboring values are not of the same color and standard denoising methods cannot be used. In this paper, we propose a new method for efficient raw data denoising that is based on a shape-adaptive DCT (SA-DCT), which was originally proposed for non-CFA data. Our method consists of three steps: a luminance transformation of the Bayer data, determining an adequate neighborhood for denoising and hard thresholding in the SA-DCT domain. The SA-DCT is applied on realistic CFA data and accounts for the signal-dependent noise characteristic using a locally adaptive threshold and signal-dependent weights. We quantitatively evaluate the method using realistically simulated test sequences reflecting typical challenges in denoising natural images and compare the results visually using real camera data. Our method is compared to the state-of-the-art methods and achieves similar performance in terms of PSNR. The visual comparison shows that our method can reach more pleasant results compared the state-of-the-art methods in terms of visual quality, while the computational complexity is kept low.

Index Terms— Denoising, raw data, color filter array, CFA data, implementation cost

1. INTRODUCTION

While Denoising is an extensively studied task in signal processing research, most denoising methods are designed and evaluated using readily processed image data, e. g. the well-known Kodak data set [1]. The noise model is usually a additive white Gaussian noise (AWGN). This kind of test data does not correspond to nowadays real-world image data taken with a digital camera.

To understand the difference, lets review the color image capture via a digital camera, which is the usual way of image capture nowadays. One pixel captures the light intensity, thus the sensor data correspond linearly to the lightness at the pixel

position. To capture color data, a color filter array (CFA) is used, which covers the pixels with a filter layer. Thus the output of the sensor is a value that represents the light intensity for one color band at one pixel position. This data cannot be displayed before further steps are applied. These steps are the white balance, the demosaicking, which leads to a full color image, and the color transformations, which adapt the linear data, which is linear to the lightness, to displayable monitor data, which is adapted to the monitor gamma and color space. These steps lead to a noise characteristic the is fundamentally different from the usually assumed AWGN: through demosaicking it is spatially and chromatically correlated and through the nonlinear color transformations the noise distribution is unknown.

As this noise characteristic cannot be easily incorporated into the denoising methods, we propose to apply the denoising to the raw CFA data – the mosaicked data linear to the light intensity with uncorrupted noise characteristics. In the raw data we observe noise with a known distribution and a signal-dependent variance, which can be precisely estimated based on measurements [2]. Up until now, despite the richness of denoising methods, denoising color image raw data has been less studied. Hirakawa extended a wavelet-based method to CFA data [3]. Zhang proposed a principle component analysis (PCA) based solution [4]. The state of the art method in image denoising, BM3D [5], was combined with a noise estimation algorithm and used for image denoising and declipping [6]. The latter gives very good results that, however, come with a high computational cost. Nonlocal methods inherently have high memory cost, as the hole image must be present in the internal memory. With high resolution image sensors this can be a limiting aspect. In this paper we therefore propose a local method for raw data image denoising, which builds on a shape-adaptive DCT [7]. The method relies on a homogeneous neighborhood for each pixel, and then a thresholding operation eliminates the noisy DCT coefficients. The neighborhood estimation prevents from oversmoothing, which is the prevailing denoising drawback. Foi et.al. adapted the method to signal-dependent noise [8], thus it can be easily used for the noise in raw data. However, the method is still not adapted to linear and mosaicked data. In this paper we therefore propose to adapt and extend the method. We calcu-

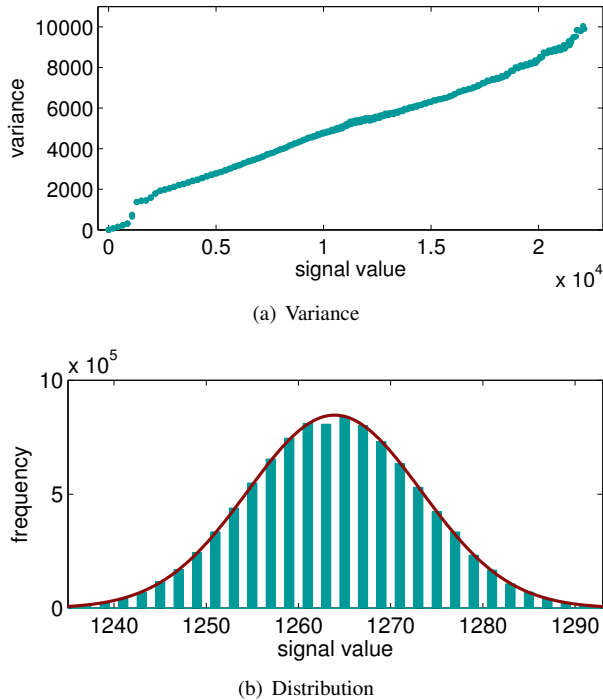


Fig. 1. Variance and distribution of the noise in the raw domain (signal values in 16 bit precision).

late the neighborhood estimation on luminance data and we propose a luminance transformation that can be directly applied to the CFA data. Additionally, we show how to adapt the SA-DCT to Bayer data, as this is the most usual CFA, and describe how the real noise characteristics from a digital camera can be obtained and included in the method. We both compare our solution to Zhang [4] and [6] and evaluate the methods in terms of visual quality and computational cost.

2. CAMERA NOISE

To apply denoising on raw data, we first need a realistic model for the camera noise. Therefore, we measure the real camera noise in the raw domain based on a series of exposures and calculate the noise variance using the photon transfer method [9]. While this measurement can be performed with any camera, we use the ARRI Alexa camera, as it delivers uncompressed raw data in 16 bit precision. Since the data is uncompressed, we can expect unaltered measurement results and additionally the individual camera processing steps are known for this camera [10]. Our method can equivalently be used for other cameras.

The Alexa camera has been developed for motion picture recordings in digital cinema applications. It has a CMOS sensor with a resolution of 2880×1620 . In front of the sensor, the camera has a filter pack composed of an infrared cut-off filter, an ultraviolet cut-off filter and a low pass filter to re-

duce aliasing. The color filter array (CFA), which is located between the filter pack and the sensor, is a Bayer mask.

The photon transfer method [9] uses two subsequent frames recorded at constant and homogenous lighting conditions. The noise variance is calculated as the mean of the difference between these two frames, the corresponding signal value is calculated as the mean over all the signal values in these frames. The graph in Fig. 1(a) shows the variance plotted over the respective mean signal value. The variance of the sensor noise can be approximated by a linear model, which matches the results reported in [11], where other cameras have been studied. We observe, however, one difference in the signal region around value 0.1×10^4 . The step in the variance curve is due to a special characteristic of the Alexa sensor, the dual-gain read-out technology. This means, the sensor read-out of the Alexa provides two different paths with different amplification (dual-gain read-out). The low amplified path provides the data for the signal range starting from 0.1×10^4 . The high amplified path saturates in the high signal values, but for the low signal values it provides a significantly higher signal-to-noise ratio. The read-out noise (offset of the variance curve) is reduced, thus the dual-gain technology enhances the low light performance of the camera. The two read-out paths are combined in the region around signal value 0.1×10^4 , which explains the step in the variance curve.

The distribution is very similar to a Gaussian distribution. In Fig. 1(b) the distribution at signal level 1265 is shown with the Gaussian approximation. The difference between the approximation and the measured histogram is small, thus we can well approximate the sensor noise n in the raw domain using a Gaussian distribution with signal-dependent variance.

$$n \sim \mathcal{N}(0, \sigma(x)) \quad \text{with} \quad \sigma^2(x) = m(x)x + t(x) \quad (1)$$

The variance $\sigma^2(x)$ is approximated as a piecewise linear function depending on the signal x , with the slope $m(x)$ and the intercept $t(x)$ based on the measurement data in Fig. 1(a). Because of the dual-gain read-out the values for $m(x)$ and $t(x)$ are piecewise constant.

We found a model for the camera noise in the raw data. Based on this model, we will describe in the next section the shape-adaptive DCT denoising method and how to integrate the noise model in the SA-DCT.

3. ADAPTIVE RAW DATA DENOISING

Our goal is to find an algorithm with high visual quality of the denoising results, which is additionally efficient in terms of a hardware implementation. Regarding the visual quality, a common problem of denoising algorithms is blurring of edges in the image. The shape-adaptive denoising algorithm [7] prevents this by using a homogenous neighborhood for denoising. As proposed by Foi we use the local polynomial approximation an intersection of confidence interval technique (LPA-ICI) to find an adequate neighborhood for each pixel.

Table 1. Results of the luminance estimation. PSNR of the denoising result using the respective luminance estimation method.

	Gaussian	h_L [12]	virtual	ADA	LSLCD [12]
PSNR	36,84	36,74	36,76	36,85	36,66

The method has been adapted for signal-dependent noise in [8]. However, it still cannot directly be used on Bayer data, because the neighboring pixels do not have the same color due to the Bayer pattern.

To find a way estimating the neighborhood based on Bayer data, we apply a luminance transformation. In color image denoising, a color space transformation from RGB to a luminance-chrominance color space (e.g. YCbCr) is usual. As the structural information in natural images is mostly contained in the luminance data, it is effective to perform the neighborhood estimation on the luminance channel only and used for denoising all three channels. In our case, we apply a similar strategy; we obtain an estimation of the luminance channel based on the Bayer data. In the next section, we discuss the luminance transformation.

3.1. Luminance transformation of bayer data

To find a luminance estimation based on Bayer data we tested different techniques: filtering with a fixed filter kernel, partial debayering, and a new method we call “virtual luminance”. For the partial debayering, we took the debayered green channel as luminance estimation directly, as the green channel is most dominant for the luminance. We used the camera debayering method (ADA), which can be applied by downloading the free “ARRI Raw Converter (ARC)” tool. As another low cost luminance estimation we used two different filters: a Gaussian filter kernel and a filter similar to the luminance filter by Jeon and Dubois [12]. We additionally calculated a luminance directly on the Bayer data by using the neighboring color values, which we call “virtual”, because the result gives us luminance values which are located between the pixels.

The results on our test image “city” in Tab. 1 show that the difference in terms of PSNR of the denoising result is marginal. The best value is reached by the camera debayering and Gaussian filtering. We use the Gaussian filtering for our method, as it shows one of the best results and additionally is a very simple and cost efficient method.

3.2. LPA-ICI for neighborhood estimation

Once we obtained a continuous luminance estimation we can apply the LPA-ICI method to find the dimension of the local homogenous neighborhood. The LPA-ICI method chooses a polynomial model that fits the pixel neighborhood data. The chosen scale of the model defines a shape around each pixel, in which thus no singularities or discontinuities are present.

The LPA-ICI method is applied in eight directions. In each direction θ_k a set of directional kernels g_{h,θ_k} with the

varying scale h is used to find an interval D .

$$D_{\hat{y}_{h_i,\theta_k}} = \left[\hat{y}_{h_i,\theta_k} - \Gamma \sigma_{\hat{y}_{h_i,\theta_k}}, \hat{y}_{h_i,\theta_k} + \Gamma \sigma_{\hat{y}_{h_i,\theta_k}} \right] \quad (2)$$

$\Gamma > 0$ is a tuning parameter, which adjusts the size of the interval. The standard deviation of the estimate $\sigma_{\hat{y}_{h_i,\theta_k}}$ is calculated by multiplying the standard deviation σ of the input with the norm $\|g_{h,\theta_k}\|_2$ of the used kernel: $\sigma_{\hat{y}_{h_i,\theta_k}} = \sigma \|g_{h,\theta_k}\|_2$.

$$\sigma_{\hat{y}_{h_i,\theta_k}} = \sigma \|g_{h,\theta_k}\|_2 \quad (3)$$

The standard deviation of the input, σ , is estimated by using the noisy observation, thus the raw data pixel value. This is a very simple estimate, but we found that the improvement using a better approximation is marginal.

The largest possible scale h_i is chosen using the ICI rule.

$$\mathcal{I}_{j,\theta_k} = \bigcap_{i=1}^j D_{\hat{y}_{h_i,\theta_k}} \quad (4)$$

This scale defines the shape dimension in the direction θ_k . The ICI rule is applied in all eight directions and thereby a neighborhood \tilde{U}_x^+ for the pixel position x is found.

3.3. Shape-adaptive DCT and denoising via hard thresholding

The shape of the neighborhood is now found for each pixel. Now the SA-DCT must be applied on the Bayer data to perform the denoising. Therefore the Bayer data is separated into the four sub channels, R , G_1 , G_2 and B , which each contain fourth of the total number of pixels.

For each color channel the SA-DCT is implemented as proposed by Foi [7]. A local estimate $\hat{y}_{\tilde{U}_x^+}$ is obtained by thresholding in the SA-DCT domain with the threshold parameter t_x set to

$$t_x = k_{thr} \sigma \sqrt{2 \ln |\tilde{U}_x^+| + 1} \quad (5)$$

The constant k_{thr} regulates the denoising strength. The global estimate is given by a weighted averaging of the local estimates.

$$\hat{y} = \frac{\sum_{x \in X} w_x \hat{y}_{\tilde{U}_x^+}}{\sum_{x \in X} w_x \chi_{\tilde{U}_x^+}} \quad (6)$$

with:

$$w_x = \frac{\sigma^{-2}}{(1 + N_x^{har}) |\tilde{U}_x^+|} \quad (7)$$

We showed how to apply neighborhood estimation on Bayer data discussed how this neighborhood can be used for denoising with SA-DCT hard thresholding. In the following we evaluate our method and compare it to other state-of-the-art methods.

4. EXPERIMENTS: IMAGE QUALITY VERSUS SYSTEM PERFORMANCE

We compare our method to the state-of-the-art method in denoising, BM3D [5], which was specifically adapted for raw data [6] and the PCA-based method from [4].

While we first tested our method on real camera data, we compare our method here using simulated camera video sequences, as this provides us a realistic reference. This test method was described in [2] and we use it for our data similarly: we rendered the high resolution image data, applied the sensor simulation including the optical lowpass of the camera optics. We obtain a simulated reference image and by including the camera noise added to the sensor data, we also obtain noisy images. These images are then denoised and compared. We tested the method using two different debayering methods: the ARRI debayering method, which is implemented in the camera processing tool freely available in the internet (Arri Raw Converter ¹) and Demosaicing With Directional Filtering and a posteriori Decision (DDFAPD) [13].

We wanted to test different noise levels. In a digital camera the noise is signal-dependent corresponding to the characteristic in Sec. 2. To obtain different noise levels, we change the simulated brightness of the image and subsequently process the images with a higher ASA level to obtain comparable results. This leads to a higher noise level, as the ASA operates as a gain: The higher the ASA value, the higher the amplification and thus the lower the signal-to-noise-ratio. Three different noise levels were simulated: ASA 800, which means a low noise level, ASA1600 and ASA 3200, which corresponds to a quite high noise level.

Table 2. Denoising Results for the noise level ASA 800.

ASA800	Debayering	PSNR	PSNRHVS	SSIM	VIF
proposed	ADA	40,390	36,766	0,993	0,777
proposed	[12]	40,703	36,777	0,993	0,810
ClipFoi	ADA	38,997	34,552	0,991	0,745
ClipFoi	[12]	39,312	34,449	0,991	0,765
Zhang	ADA	40,815	37,183	0,994	0,775
Zhang	[12]	41,325	37,265	0,994	0,802

4.1. Visual quality of denoising results

First, we want to compare the method quantitatively, thus we calculate quality metrics enabling us to do so. We used PSNR,

¹http://www.arri.com/camera/digital_cameras/tools/arriraw_converter/

Table 3. Denoising Results for the noise level ASA 1600.

ASA1600	Debayering	PSNR	PSNRHVS	SSIM	VIF
proposed	ADA	38,395	34,537	0,988	0,696
proposed	[12]	38,682	34,787	0,988	0,737
ClipFoi	ADA	40,649	37,058	0,995	0,792
ClipFoi	[12]	41,233	36,794	0,994	0,818
Zhang	ADA	38,957	34,821	0,990	0,676
Zhang	[12]	39,458	35,089	0,990	0,703

Table 4. Denoising Results for the noise level ASA 3200.

ASA3200	Debayering	PSNR	PSNRHVS	SSIM	VIF
proposed	ADA	35,890	31,781	0,976	0,586
proposed	[12]	36,120	32,239	0,979	0,631
ClipFoi	ADA	39,272	35,458	0,991	0,726
ClipFoi	[12]	39,690	35,237	0,991	0,752
Zhang	ADA	36,342	31,678	0,980	0,546
Zhang	[12]	36,726	32,031	0,981	0,567

as it is very usual, and additionally we used three metrics that according to [2] correlate better with the human perception: a PSNR adapted to the human visual system (PSNR-HVS) [14], structural similarity index (SSIM)[15] and the visual information fidelity (VIF) [16]. While the Bayer SA-DCT reaches the highest VIF for the 800ASA sequences. For higher ASA, thus higher noise levels, the ClipFoi[6] method reaches the highest values, while SA-DCT and PCA[4] show the same results. We conclude that our method achieves competitive results with respect to usual quality metrics.

When comparing the results visually we think our method provides slightly better results. We analyzed the results and found that this is due to the low correlation, the remaining denoising error appears less disturbing as it is less correlated. The spatial correlation of the denoising error – also called “method noise” – is calculated and shown in Fig. 3(c).

However, as the visual perception is not well represented by the metrics, we can evaluate the results by comparing the difference images $I_{diff} = I_{ref} - I_{denoised}$. In the review of denoising algorithms in [17] the difference images were used to compare the algorithms. An ideal difference image would look like noise, when image structure comes through this means usually that image details is lost during denoising. Fig. 2(a) shows the difference of the reference to the image denoised by Zhang [4]: The image structure is clearly visible, which indicates that the image was blurred. In Fig. 2(b) we observe the finer image structures, thus only finer details seem so get lost. The difference image is colored, which means that the error is an offset and thus the color of the denoised image does not correspond to the correct color of the reference. The difference image in Fig.2(c) shows the difference for the proposed Bayer SA-DCT. The image is most noisy looking, and thus not much image structure is lost and no color shift must be expected.

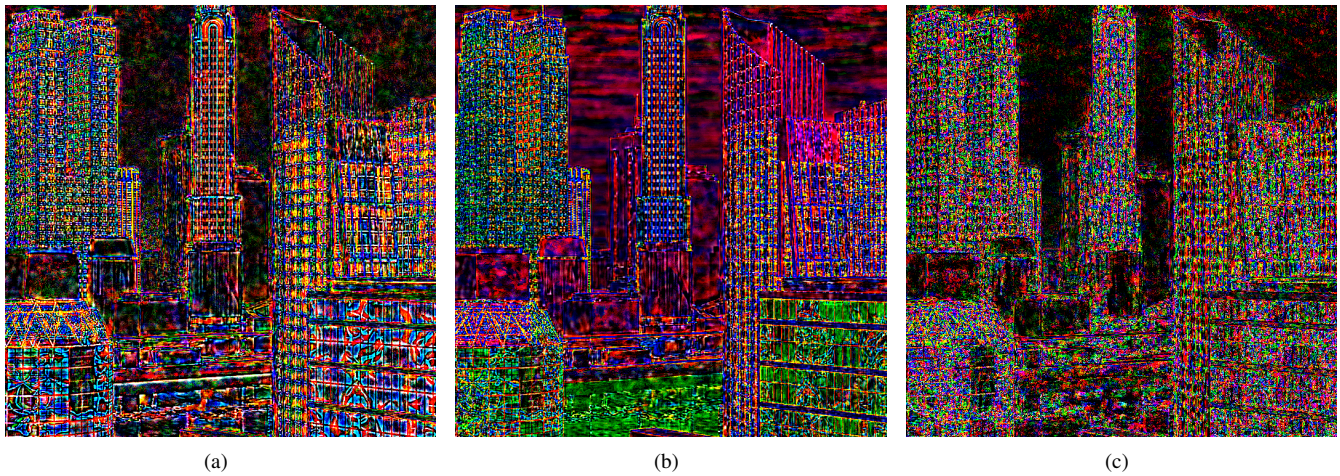


Fig. 2. Difference to the reference for the test image “City” denoised using (a) the algorithm of Zhang [4], (b) the Foi [6] and (c) the proposed version of the SA-DCT.

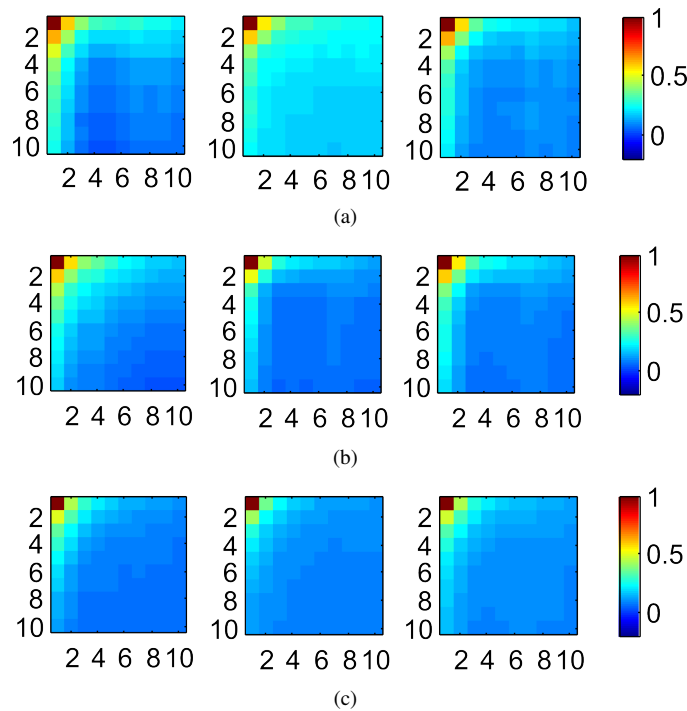


Fig. 3. Spatial correlation of the difference images for red (left), green (middle) and blue (right) channel. Displayed is the correlation coefficient to its neighboring pixels in a 10×10 neighborhood. From top to bottom: Foi [6], Zhang [4] and the proposed Bayer Data SA-DCT. Ideally only noise should be removed from the image and thus the correlation of the difference image should be low.

4.2. Implementation aspects

Denosing algorithms as BM3D usually assume that the complete noisy image is available at each pixel position. In stream based image processing, as it is usual in embedded systems and communication, only one pixel at a time is available. In these applications an additional buffer must be implemented to provide the neighborhood for the denosing step. For algorithms like BM3D therefore the memory cost is quite high. Our method requires only a local neighborhood for the denosing step and is therefore better suited for stream based processing. Additionally it operates on Bayer data and the raw Bayer data has only one value per pixel whereas the processed RGB data has three values per pixel. That means a three times lower complexity can be expected compared to algorithms that require the fully processed image data.

5. CONCLUSION

We proposed a method for real camera Bayer data denosing based on a neighborhood estimation combined with a shape-adaptive DCT. While the method has been proposed for RGB or grayscale image data, it couldn't be applied to Bayer data directly. To perform the SA-DCT on Bayer data we propose a LPA-ICI based neighborhood estimation on the luminance data. As the luminance data is not available in the Bayer data, we estimate the luminance efficiently using different methods. The best tradeoff between computational cost and quality was found using Gaussian filtering. Based on the neighborhood estimation a hard thresholding is performed on the coefficient of the shape-adaptive DCT. The threshold includes the noise variance and we show how a real camera noise characteristic can be integrated. To evaluate our method we compare it with two algorithms: a PCA-based CFA denosing and a BM3D-based denosing that uses noise variance estimation. Our method achieves competitive results in terms of PSNR. We determine the visual quality and show that our method can lead to better visual quality than other methods, while the computational cost is reduced.

6. REFERENCES

- [1] "<http://r0k.us/graphics/kodak/>," .
- [2] Tamara Seybold, Christian Keimel, Marion Knopp, and Walter Stechele, "Towards an evaluation of denosing algorithms with respect to realistic camera noise," *IEEE International Symposium on Multimedia*, vol. 4, no. 5, 2013.
- [3] K. Hirakawa, X.-L. Meng, and P.J. Wolfe, "A framework for wavelet-based analysis and processing of color filter array images with applications to denosing and demosaicing," in *ICASSP*, 2007, vol. 1, pp. I-597 – I-600.
- [4] L. Zhang, R. Lukac, X. Wu, and D. Zhang, "PCA-Based spatially adaptive denosing of CFA images for single-sensor digital cameras," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 797 – 812, 2009.
- [5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denosing by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080 – 2095, 2007.
- [6] Alessandro Foi, "Practical denosing of clipped or overexposed noisy images," in *Proc. 16th Eur. Signal Process. Conf., EUSIPCO 2008*, 2008.
- [7] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denosing and deblocking of grayscale and color images," *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1395 – 1411, may 2007.
- [8] A. Foi, V. Katkovnik, and K. Egiazarian, "Signal-dependent noise removal in pointwise shape-adaptive DCT domain with locally adaptive variance," in *EU-SIPCO*, 2007.
- [9] "EMVA 1288, standard for characterization of image sensors and cameras," 2010.
- [10] Stefano Andriani, Harald Brendel, Tamara Seybold, and Joseph Goldstone, "Beyond the kodak image set: A new reference set of color image sequences," in *ICIP*, 2013, pp. 2289–2293.
- [11] H. J. Trussell and R. Zhang, "The dominance of poisson noise in color digital cameras," *ICIP*, pp. 329 – 332, 2012.
- [12] G. Jeon and E. Dubois, "Demosaicking of noisy bayer-sampled color images with least-squares luma-chroma demultiplexing and noise level estimation," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 146–156, 2013.
- [13] D. Menon, S. Andriani, and G. Calvagno, "Demosaicing with directional filtering and a posteriori decision," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 132–141, Jan. 2007.
- [14] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *VPQM*, 2006.
- [15] Z. Wang, A. C Bovik, H. R. Sheikh, and E. P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600 – 612, 2004.

- [16] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [17] A. Buades, B. Coll, and J. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490 – 530, 2005.

Removal of Circular Edge Noise of Retinal Fundus Images

A. Ektesabi¹ and A. Kapoor¹

¹Faculty of Science, Engineering and Technology, Swinburne University of Technology, Victoria, Australia

Abstract - Many studies have been conducted in the field of ophthalmology, resulting in highly accurate and fast detection of retinal features. However, when considering a generalized method which can be implemented on all images, with different sizes and resolution, some of these techniques may result in an inaccurate or invalid localization. To overcome this problem, it is important to reduce all possible noises in the preliminary stage of analysis. This includes the illuminated light which may occur in the edges of the images. In this study, circular approximation has been used to create a trimming circle in which these bright regions are removed. The obtained result, indicates, significant improvement on Optic Disk localization, which previously resulted in inadequate outcome.

Keywords: Noise, Fundus Image, Trimming Circle, Image Processing, Ophthalmology

1 Introduction

Since the 1990s despite the growth in the elderly population, the overall visual impairment has decreased significantly. These were mainly due to improvements in disease related studies and technologies, as well as an increase concern in regard to public health, and availability of health care services [1].

Table 1: World Health Organization (WHO) estimation of visual impairment [1-2]

Year	Population (Billion)	Blind (Million)	Low vision (Million)	Visually impaired (Million)
1975	4.1	28		
1996	5.8	45	135	180
1999	Introduction of 2020 Vision by WHO			
2002	6.3	37	124	161
2006	6.6	45	265	314
2010	6.9	32	246	289
2020	7.9	76	Estimated in 1990s	

Based on the studies conducted by World Health Organization (WHO), currently there are about 285 million people who are visually impaired, from which 39 million are blind and 246 million have low vision, which consists of moderate to severe visual impairment. Over the years, the leading causes of blindness consisted diseases such as

cataract, glaucoma and age related macular degeneration. It has also been indicated that 80% of visual impairments in the world can be avoided if detected early [1].

Due to great interest in health and advancements in biomedical technologies, the ophthalmic practices, has been significantly influenced. Different eye diseases are now studied, diagnosed, treated and monitored in depth as the result of rapid enhancements in image capturing and analysis. The surgeons and ophthalmologist now, have better understanding of the problems which the patients might face and can act in a more precise but timely manner.

However, despite these significant reductions in complications and improvements in technology, there are still many patients in the developing countries who are in need of fast but affordable diagnosis and treatment procedures.

Therefore it is essential to continue developing new methodologies and techniques which are accurate, reliable, inexpensive and accessible for early diagnosis of diseases.

2 Image Processing

Varieties of the image processing processes follow main steps of [3]:

1. Image Acquisition
2. Image Manipulation
 - Segmentation
 - Normalization
3. Feature Localization
4. Feature Extraction
 - Matching

The first main step, commonly known as image acquisition is very important as it directly affects the accuracy and precision of the overall process. While collecting these data, it is important to consider the available space to save these data as well the required processing time of the application. In cases where fast processing time but not great precision is required, the very high quality image may not be necessary. Hence it is important to consider factors such as, accuracy, processing time, data storage facilities and resources available to capture the data, in order to collect the most feasible and logical images.

In the manipulation stage, usually the RGB colored images are changed to the grey scaled imaged. This decreases the amount of data and increases the speed of the processing. In this stage, other preliminary procedures such as filtering and contrast adjustment are also implemented.

Feature localization is the step in which key regions of interest are identified and located.

In the last stage, feature extraction, the information obtained from the localized regions is studied and the objectives of the study are fulfilled.

The flow chart below, suggest the steps undertaken in the studies of retinal image.

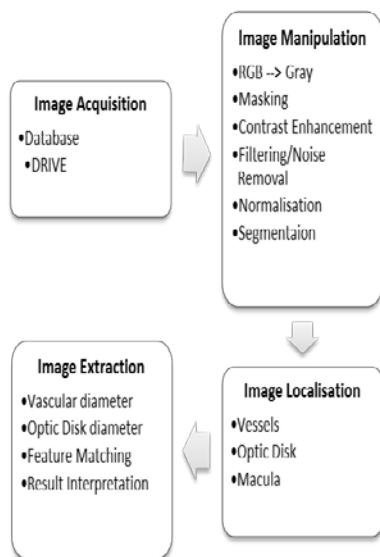


Fig. 1. Flow chart of the image processing steps

For the purpose of this study, the use of retinal images is considered. The images used are from DRIVE database [4]. The collection of these images was initiated by Staal et.al. This open source database consists of twenty colored retinal images, captured by Canon CR5 non-mydratic 3CCD camera with a 45° field of view. These images which are available digitized to 768×584 pixels, 8 bits per color channel [4].

2.1 Retinal Fundus Image Analysis

Over the past few decades, there have been many advances in retinal image studies. Majority of the studies conducted were concentrating on localizing the Optic disc (OD) [5-9] and detection of blood vessels [10-13], which mainly focused on the last two stages of the image processing.

However, when it comes to fully automated detection of these features, the accuracy of detection varies significantly. This may have been due to the wide range of images, which

are collected from wide range of instrumentations and or the effect of responses from individual patients.

As a result, the second stage, initial preliminary manipulation of the image is crucial in automation of image processing.

3 Suggested Method

Majority of the studies conducted on retinal images, concentrate on the image localization and segmentation steps.

However, based on the visual studies conducted on the obtained results, there are times were the accuracy of the results are affected by the presence of noise, which considered as illuminated bright spots, on the edges of the retinal image [14]. Presences of these bright regions maybe have been due to the eye response or the refraction of the light within the eye. An example of which can be seen in figure 2.

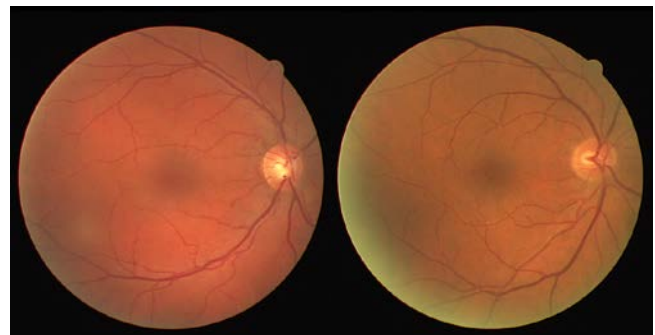


Fig. 2. Comparing two different retinal images, it can clearly be seen that in the left hand corner of the right hand image, there is large bright region. However, in the left hand image, no such regions are viewed.

In these instances, the detection of the bright features, such as the OD, may have been disturbed, resulting in wrong localization of the OD.

Many have attempted to reduce these noises, via filtering or adjusting the contrast of the image. However, as result, some information may have been lost and the accuracy of the detection directly affected.

In this study, we have proposed an adaptive trimming circle which may be used to remove those regions.

3.1 Theoretical Concept

The suggested trimming circle, is approximated by calculating the radius of the circle, and has been used to reduce the noise in the obtained fundus images.

Previously, a study conducted by Zhang et al suggested a trimming boundary [15], based on the equation:

$$X^2 + Y^2 + AX + BY + C = 0 \quad (1)$$

$$C_x = -\frac{A}{2} \quad (2)$$

$$C_y = -\frac{B}{2} \quad (3)$$

$$r = \sqrt{\frac{A^2 + B^2}{4 - C}} \quad (4)$$

However, in this study the following is considered. We know that a circle, with the center (a, b) , is represented by the equation:

$$(x - a)^2 + (y - b)^2 = r^2 \quad (5)$$

Expanding this equation would result in:

$$x^2 + y^2 - 2ax - 2by + a^2 + b^2 - r^2 = 0 \quad (6)$$

Equating (1) with (6) would provide:

$$X^2 = x^2 \Rightarrow X = x \quad (7)$$

$$Y^2 = y^2 \Rightarrow Y = y \quad (8)$$

Moreover,

$$AX = -2ax$$

We know that $X = x$, therefore:

$$A = -2a \Rightarrow a = -\frac{A}{2} \equiv C_x \quad (9)$$

Similarly:

$$BY = -2by$$

We know that $Y = y$, therefore:

$$B = -2b \Rightarrow b = -\frac{B}{2} \equiv C_y \quad (10)$$

$$C = a^2 + b^2 - r^2 \quad (11)$$

Substituting (9) and (10) into (11):

$$C = \frac{A^2}{4} + \frac{B^2}{4} - r^2 \quad (12)$$

Making r the subject:

$$r = \sqrt{\frac{A^2 + B^2 - 4C}{4}} \quad (13)$$

This is then applied to all the images in the database.

3.2 Implementation

Based on the previous section, it can be said that in order to implement the suggested theoretical concept, the radius and the location of the center needs to be estimated.

In order to estimate the center, it is suggested to first mask the image, using the previously suggested technique in [16].

Using the obtained mask, the first and last white pixels across the columns of the image are detected. The center is then estimated by finding the middle value between these two pixels. This is shown as red crosses in the next figure.

With the aid of the estimated center, we then find the first and last white pixels, along the rows of the image. To define a more accurate center, the average values of the two red crosses are used to redefine the position of the center, marked as a blue cross in the figure 3.

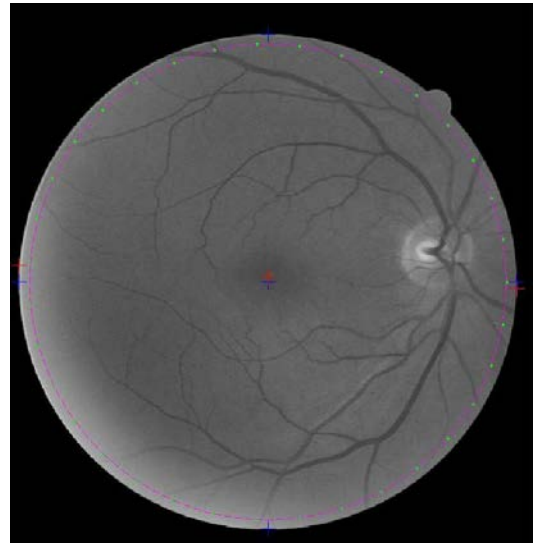


Fig. 3. Suggested trimming circle, the initial step is marked as red '+', which is then followed by blue '+'. The purple markings define the estimated trimming circle, while the green color defines the estimated ellipse.

The radius can then be calculated using the obtained center value. The radius is marked in purple in the image.

Since the images may not have been completely circular and may have had an elliptical shape, the radius in horizontal axis (short axis) and vertical axis (long axis) have been calculated separately and the ellipse is formed and drawn in

the image, which is can be seen as the green line in the figure 3.

4 Results and Discussion

As it can be seen in figure 3 and reviewing and analyzing all the images in the database, it can be suggested that the circular trimming approximation is sufficient and can be used instead of the elliptical approximation.

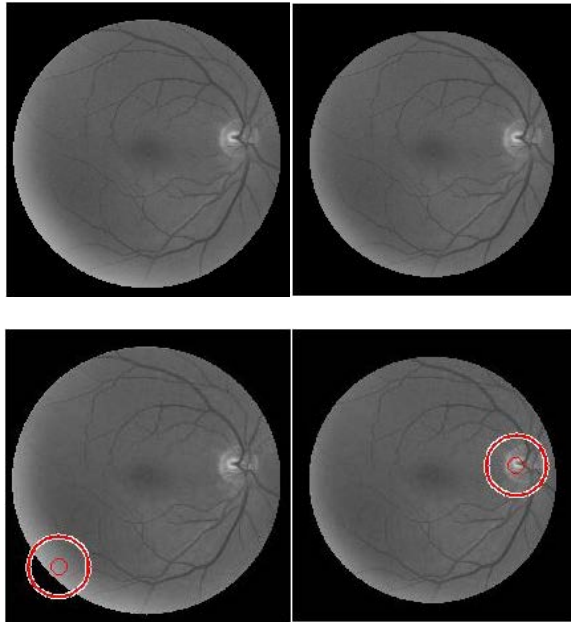


Fig. 4. The top left image represents the original gray scaled image, and the red circle on the bottom left image represents its located OD. The top right hand image indicates the retinal image which has suggested trimming circle applied to it. The red circle on the right hand bottom image represents its OD.

The suggested method has been implemented on all the images in the database. The results obtained indicate that localization of the key features, in this instance the OD has improved significantly.

5 Conclusions

While analyzing retinal images, there are times, were the result is affected by the illuminated light at the edges of the image. To avoid or minimize this effect a circular trimming circle is suggested to remove that noise. The obtained results proved to be promising and have resulted in a more accurate and robust detection of features such as the OD.

6 References

- [1] World Health Organisation. (March 2014). *Global data on visual impairments 2010*. Available: <http://www.who.int/blindness/GLOBALDATAFINALforweb.pdf?ua=1>
- [2] GEOHIVE. (April 2014). *Population of the entire world, yearly, 1950-2100*. Available: http://www.geohive.com/earth/his_history3.aspx
- [3] H. Mehrabian and P. Hashemi-Tari, "Pupil Boundry Detection for Iris Recognition Using Graph Cuts," in *Proceedings of Image and Vision Computing New Zealand 2007*, Hamilton, New Zealand, 2007, pp. 77-82.
- [4] J.J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, B. van Ginneken, "Ridge based vessel segmentation in color images of the retina", *IEEE Transactions on Medical Imaging*, 2004, vol. 23, pp. 501-509.
- [5] A. Youssif, A. Ghalwash and A. Ghoneim, "Optic disc detection from normalized digital fundus images by means of a vessels direction matched filter". *IEEE Trans Med Imag*, 2008, vol. 27, pp.1118.
- [6] R. Rangayyan, X. Zhu, F. Ayres, A. Ells, "Detection of the optic nerve head in fundus images of the retina with Gabor filters and phase portrait analysis", *Journal of Digital Imaging*, 2010, vol.23, pp. 438453.
- [7] X. Zhu, R. Rangayyan, A. Ells, "Detection of the optic nerve head in fundus images of the retina using the hough transform for circles", *Journal of Digital Imaging*, 2010, vol.23, pp.332341.
- [8] S. Sekhar, W. Al-Nuaimy and A. Nandi, Automatic localization of optic disc and fovea in retinal fundus, 16th European Signal Processing Conference, 2008
- [9] D. Welfer, J. Scharcanski, C. Kitamura, M. Dal Pizzol, L. Ludwig and D. Marinho, Segmentation of the optic disk in color eye fundus images using an adaptive morphological approach, *Comput. Biol. Med*, 2010, vol. 40, pp.124137
- [10] X. Xu, M.D. Abr`amoff, G. Bertelsen, and J.M. Reinhardt (2012), "Retinal Vessel Width Measurement at Branching Points using An Improved Electric Field Theory-Based Graph Approach", *Medical Imaging*, 2012
- [11] M.M. Fraz, S.A. Barman, P. Remagnino, A. Hoppe, A. Basit, b. Uyyanonvara, A.R. Rudnicka, C.G. Owen, "An approach to localize the retinal blood vessels using bit planes and centerline detection", *Computer Methods and Programs in Biomedicine*, 2012, pp. 600–616
- [12] Y. Yamamoto, Y. Yamamoto, A. Marugame, M. Ogura, A. Saito, K. Ohta, M. Fukumoto, T. Murata, " Age-Related Decrease of Retinal Vasculature Are Identified with a novel Computer-Aided Analysis System", *Tohoku J. Exp. Med*. 2012, vol. 228, pp. 229-237

- [13] S.C. Cheng and Y.M. Huang, "A Novel Approach to Diagnose Diabetes Based on the Fractal Characteristics of Retinal Images", *IEEE Transactions on Information Technology in Biomedicine*, 2003, Vol. 7, pp.163-170
- [14] I. Jamal, M. Akram, A. Tariq, "Retinal Image Preprocessing : Background and Noise Segmentation", in *TELKOMNIKA*, 2012, Vol, 10, No.3, pp. 537-544.
- [15] Z. Zhang, F.S. Yin, J. Liu, W.K. Wong, N.M. Tan, B.H. Lee, J. Cheng, T.Y. Wong, "ORIGA-light : An Online Retinal Fundus Image Database for Glaucoma Analysis and Research", *32nd Annual International Conference of the IEEE EMBS*, 2010, 3065-3068, Argentina.
- [16] A. Ektesabi and A. Kapoor, "Exact Pupil and Iris Boundary Detection", *International Conference on Control, Instrumentation, and Automation (ICCIA)*, 2011, 2:1217-1221, Shiraz.

A New Approach for Removing Haze from Images

Vinuchackravarthy Senthamilarasu¹, Anusha Baskaran², and Krishnan Kutty²

¹ Centre for Research in Engineering Sciences and Technology (CREST),
KPIT Technologies Ltd, Pune, Maharashtra, India.

Abstract-The presence of suspended particles like haze, fog, mist, smoke and dust in the atmosphere deteriorates quality of captured image. It is of paramount importance to reduce these deteriorating effects from the image for various image based applications; viz. ADAS, CCTV surveillance, etc. In this paper, this interesting problem of enhancing the perceptual visibility of an image that is degraded by atmospheric haze is addressed. An efficient way of estimating the transmission map and the atmospheric light is proposed, which is further used in reducing effects of haze from the image. The underlying idea is to restore the true color of each pixel by using our proposed method that minimizes the lowest of RGB values per pixel. This is accomplished using the HSV color space and the haze image model. In comparison with the other state of the art methods that are available in literature, the proposed method is shown to be capable of recovering better haze-free images both in terms of visual perception and quantitative evaluation.

Keywords: HSV color space, Atmospheric light, Transmission map, De-hazing

1. INTRODUCTION

There has been tremendous research in the areas related to Advanced Driver Assistance Systems (ADAS), video surveillance systems etc., in the last few years. As the need for such systems increase, the associated challenges also increase. These challenges can arise due to technology (footprint on embedded platform, cost etc.), or because of nature (weather, ambient light etc.). In the field of image processing and computer vision, image degradation due to the natural factors manifests as a very challenging problem. There are many aspects that affect the quality of an image in terms of visual perception and interpretation. Some aspects caused due to natural factors include loss of contrast, poor rendering of color and loss of depth information. When such an image is to be processed, it manifests into the reduced image understanding and difficulty in feature detection and identification of the object of interest. To overcome these disadvantages, many researchers have worked in areas related to the removal of atmospheric effects like haze, fog, smoke etc. It is a well-known phenomenon that every particle of significant size in the atmosphere scatters and absorbs light from the scene; thereby causing degradation in the scene visibility. This degradation in acquired images caused due to homogeneous atmospheric haze is modelled as:

$$I(x) = D(x)T(x) + A(1 - T(x)) \quad (1)$$

The first term $D(x)T(x)$ represents decayed scene radiation and hence is called as the direct attenuation term. Here, $D(x)$ denotes haze free image intensity value at pixel x and $T(x)$ denotes the transmission map describing the amount of the light that is not scattered. It is to be noted that transmission map has direct relation to the depth of the scene point from the camera. The second term $A(1-T(x))$ represents the scattered light from the atmospheric particles. It is called the air light attenuation term in which A describes the global atmospheric light and is independent of the position of object point. To restore visibility of images from given hazy image $I(x)$, one needs to infer global atmospheric light A and transmission map $T(x)$ from the given image information.

Currently, many image de-hazing algorithms are available in the literature. These image restoration and enhancement algorithms are either multiple images [2, 3] based, or single image [5-13] based techniques. Although the multiple images based de-hazing algorithms are efficient in image restoration, they rely heavily on clear scene information which makes them unsuitable for real time systems. Thus, single image based de-hazing algorithms are preferred. Of late, single image de-hazing algorithms have progressed significantly. These approaches can be broadly classified into image enhancement [4] based methods and physical model [5-13] based methods. It is to be noted here that the model based de-hazing algorithms restore images with very minimal loss of information. However, the only challenge of these algorithms is that they require real world information viz. Depth and global atmospheric light of the scene. Literature available on de-weathering is significant in terms of model based de-hazing techniques. Narasimhan et al. [5] proposed a user interactive algorithm to remove haze effect from the single image. In their algorithm, the depth map is computed by calculating the pixel distance from the user approximated vanishing point. The drawback of this method is the need for human intervention. Tan [6] proposed a method to automate the single image de-hazing by rewriting the haze removal model and by developing a cost function using the Markov random field framework. Using the white balanced hazy image, he increased the contrast of an input image without considering the atmospheric light. In Fattal's work [7], statistically uncorrelated shading and transmission field are separated and utilized to de-haze the thin hazy images. While using the implicit graphical model, he extrapolated the solutions to pixels with unreliable transmission for de-hazing. Guo et al. [8], in their work,

transformed the hazy image into YCbCr space and used retinex theory on the computed luminance component to de-haze single images using physical model. He et al. [9] proposed a dark channel prior which assumes that color channel of each pixel having very low intensity gets more contribution from the atmospheric light. In their method, the soft matted transmission map are estimated directly from the dark channel prior and utilized for de-hazing images at higher computational cost. Zhang et al. [10] proposed an improved optical model by classifying the transmission into objective and distance transmission. They used color clustering technique to segment different objects and thereby estimating the depth and atmospheric light depending on the position of object point. A fast method of enhancing the visibility of hazy image has been proposed by Tarel et al [11] for both color and gray value images. The proposed method computes atmospheric veil instead of inferring depth map for de-hazing and uses median of median filter for preserving edges with large depth jumps. Gibson et al. [12] presented a two-step single image de-hazing method using adaptive Wiener filter. In the first step, they computed naïve statistical estimates using foggy image and then updated the estimates in the second step using naive defogged image to efficiently smooth the transmission map. Lan et al. [13] showed that efficiency of de-hazing algorithm improves when the hazy images are preprocessed prior by removing sensor blur and noise.

In the present work, a simple yet effective de-hazing algorithm using HSV color space has been proposed. Fundamentally, the algorithm minimizes the RGB channel of each pixel by using the HSV values and the global atmospheric light to bring back the true color. The process of restoring color begins with normalizing the RGB values of hazy image with optimized global atmospheric light. This is followed by transformation of the normalized RGB space into HSV color space. The computed saturation (S) and intensity values (V) of each pixel are utilized for estimating the transmission map. The computed transmission map is further processed and smoothed using the Guided Image filter [14] and is used with our modified de-hazing model for restoration of hazy image. The restoration results using our proposed approach has enhanced perceptual visibility and are subjectively better when compared with results from He et al. [9], Tarel [11] and Gibson [12]. We have also conducted a quantitative comparative study of our results with respect to [9], [11] and [12], by using the visible edges segmentation method proposed by Hautiere et al. [15].

The rest of this paper is organized as follows: In section I, an effective way of computing the global atmospheric light and normalizing the RBG channels of input hazy image is detailed. In Section II, we describe the estimation of transmission map from the image using the HSV color space. In Section III, the process of smoothing of transmission map using guided filter is described. The de-hazing of hazy images using the modified physical model is demonstrated in the Section IV. In Section V, the results of the proposed algorithm and its comparison

with other well-known algorithms are shown. Section VI concludes the paper with discussion and summary.

2. IMAGE NORMALIZATION

We intend to normalize the RGB channels of input hazy image using the computed global atmospheric light. Global atmospheric light (A) is a position independent scalar value that represents the atmospheric light of the scene. It plays an important role in changing the vivacity of the image. It is, therefore, vital to compute A precisely to restore better visibility in a hazy image. In [9], it is seen that the global atmospheric light is calculated from the brightest pixels in hazy image. However, we have found out that the presence of sky or saturated region would exaggerate the value of A . Applying this over-estimated A in any de-hazing model tends to increase contrast. In some cases, it also creates false color in the de-hazed output. In order to avoid any saturated or sky region during the calculation of A , our method initially segments out the sky and other saturated regions in the image. This is achieved by exploiting the obvious condition that pixels having saturated or sky regions will have maximum intensity values in all three RGB channels. In line with this, we present a simple thresholding technique to segment out the sky or saturated region in the image. The method initially normalizes each channel (in RGB color space) of the hazy image with their global maximum intensity value and thereafter converts them ($J_c(x)$) into binary image ($B_c(x)$) with 95% threshold (See Equation 2 and 3). The intersection of the three binarized channels determines the exact sky or saturated regions. The working of this technique is as shown in Figure 1.

$$J_c(x) = \frac{I_c(x)}{\max(I_c(x))} \quad (2)$$

$$B_c(x) = \text{graytobinary}(J_c(x), 0.95) \quad c \in (R, G, B) \quad (3)$$

$$\text{Saturated_region} = B_R(x) \cap B_G(x) \cap B_B(x) \quad (4)$$

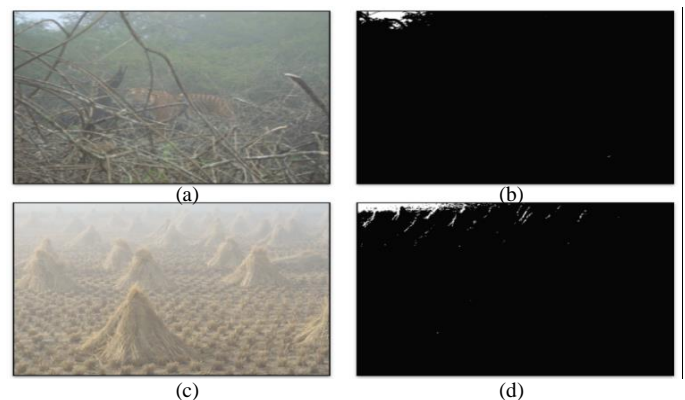


Figure 1. Input hazy images (a & c) and binary images highlighting the segmented sky region (b & d)

The proposed segmentation method helps masking the detected sky or saturated region prior to the calculation of atmospheric light. By using this masked image, the atmospheric light for each channel are calculated by averaging 0.1% of the remaining brightest pixels' intensity values. Thus the computed atmospheric light values pertaining to R, G and B channels is averaged to obtain the global atmospheric light A . This simple and effective approach of calculating global atmospheric lighting avoids overdoing of de-hazing model and hence reduces false color and prevents excessive contrast to the restored image. The global atmospheric light for the input hazy images given in Fig 1(a) and 1(c) are estimated to be 220 and 221 respectively. The estimated global atmospheric light is used to normalize the RGB channels of the input hazy image $I_c(x)$:

$$I'(x) = \frac{I_c(x)}{A}, c \in (R, G, B) \quad (5)$$

3. ESTIMATING THE TRANSMISSION MAP

A new and much simpler method of computing transmission map is presented in this paper. Transmission map implies the amount of light transmitted through haze from the object point to the camera and hence it is inversely related to depth map. For an object at a far distance from the camera, the transmission value will be lesser; while for a closer object, the transmission value will be closer to 1. In order to make the computation of transmission map easy and effective, we propose a new way of computing the transmission map using HSV color space. In the RGB space, a 'true color' is defined as a color in which one of R, G and B values is either zero or very close to zero. Our approach is to restore the closest true color to every pixel in the image that is affected by haze. First, the method transforms the normalized image $I'(x)$ into its HSV color space and then uses the computed saturation ($S(x)$) and intensity values ($V(x)$) of the transformed image to compute the transmission map as given in equation (6):

$$T(x) = 1 - q \cdot V(x) \otimes (1 - S(x)) \quad (6)$$

' \otimes ' indicates pixel to pixel multiplication.

Where ' q ' is the multiplication factor which influences the quantum of haze to be removed from the input image. For the present study, q is assumed to be 0.95. In order to avoid the transmission map in the sky or saturated region exceeding unity, the transmission values in those regions are made to 0.95. A low level smoothing is carried out over the computed transmission map to smooth out high variations between neighboring pixels. It is done by dividing the transmission map into patches and assigning the minimum of transmission values in the patch to every pixel inside the patch. This low level smoothing of the transmission map is then followed by intense smoothing using the guided filter approach. For all example images presented in this paper, the patch size used for low level smoothing is a 3×3 window. Fig. 2(a) & 2(c) show the results of low level smoothing of images shown in Fig. 1(a) and 1(c). Fig 2 clearly shows the maximum transmission values in the

less hazy region and minimum transmission values in the regions far from the camera.

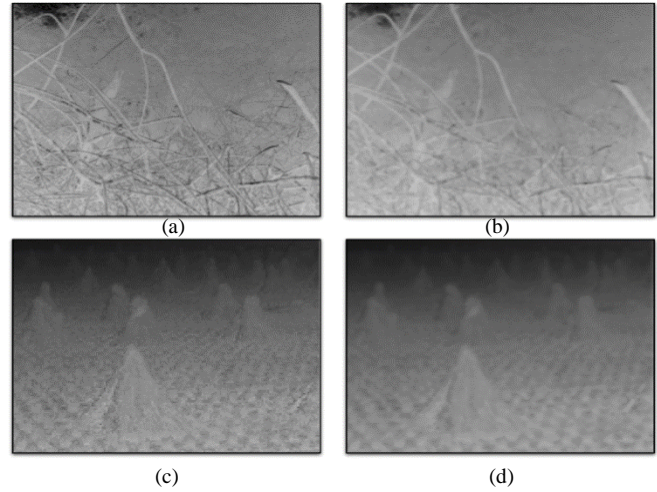


Figure 2. Transmission map for the given input images computed after low level filtering (a & c) and guided filtering technique (b & d).

4. GUIDED FILTER

He et al. [14] proposed an explicit image filtering technique called Guided Filtering technique which is effective than Bilateral filter to preserve edges while removing small fluctuations in the transmission map. We use this approach to smooth the transmission map. The guided filtering technique preserves the edges by considering local linear model between the content of the guidance image and the input image. In the present study, transmission map before ($T(x)$) and after low level smoothing ($L(x)$) are considered as the guidance image and input image respectively. Thus the smoothed transmission map $T'(x)$ is given as

$$T'(x) = a_k T(x) + b_k \quad (7)$$

Here a_k and b_k are the linear coefficient for a subset ω of transmission map, centered at pixel k . The cost function that minimizes the difference between $T'(y)$ and $L(y)$ is considered as follow:

$$E(a_k, b_k) = \sum_{y \in \omega} ((a_k T(y) + b_k - L(y))^2 + \alpha a_k^2) \quad (8)$$

E is the regularization parameter which is considered as 0.01 in the current work. The given cost function (8) is minimized for a_k and b_k and derives out to be

$$a_k = \frac{\frac{1}{|\omega|} \sum_{y \in \omega} T(y)L(y) - \mu_T \mu_L}{\sigma_T^2 + \epsilon} \quad (9)$$

$$b_k = \mu_L - a_k \mu_T \quad (10)$$

Where, μ_T and σ_T are the mean and standard deviation of $T(y)$ within the subset ω centered at pixel k and μ_L is the mean of $L(y)$ within the considered subset. Since a_k and b_k are computed for pixel k continue to vary and get updated when computed with different windows; the average value of a_k and b_k for various window are computed and is used in Equation(7) for smoothing the transmission value at pixel k . For the example images given in this paper, a subset size of 5×5 pixels is used. The guided filtered transmission map corresponding to Fig. 2(a) and 2(c) are shown in the Fig. 2(b) and 2(d).

5. MODEL BASED IMAGE DE-HAZING

The image hazing model given in Equation 2 is rewritten appropriately in the current work to incorporate the computed transmission map $T'(x)$ and normalized RGB values $I'(x)$ of hazy image for de-hazing:

$$D(x) = \frac{(I'(x) - 1 + T'(x))}{T'(x)} \quad (11)$$

This rewritten model minimizes the lowest of the RGB values for each pixel to produce true color. For the input images in Fig. 1, the resultant images obtained using the proposed de-hazing model is given in the Fig. 3.

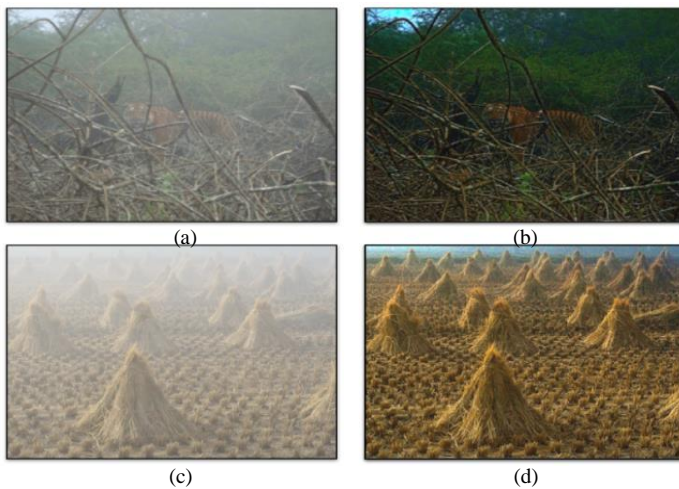


Figure 3: Resultant images corresponding to input images in Fig. 1(a) & 1(c) using the proposed algorithm

6. RESULTS AND DISCUSSIONS

The proposed algorithm was developed in MatlabTM on a 3 GB RAM, 2.40 GHz Intel CORETM i3 processor. The images and visible edge segmentation technique source code used for comparing the results are taken from the webpage of Gibson et al. [12] and Hautiere et al. [15]. The eight input hazy images used in our study are shown in Fig. 4. The entire process of removing haze from the input images is completely automatic. The de-hazed test images obtained from different algorithms along with smoothed transmission map are given in the Fig. 5-12. In comparison with the other methods, the proposed

algorithm is able to remove the varying amount of haze (i.e., mild to dense) effectively. The performance of the proposed method is quantified using the visibility edge segmentation technique indicators e , r and σ . These technique converts the input and restored images into gray scale value and compares to compute the mentioned indicators. The parameter e describes the rate of newly visible edges after restoring the visibility. The visible edges are computed using Weber contrast calculator and contrast below 5% are exempted from the calculation. Fig. 13 shows the comparison of e for different test images restored using considered algorithms. It is evident from the plot that the proposed algorithm performs exceedingly well in restoring visible edges. The descriptor r demonstrates the geometric mean ratios of visibility level in the image before and after de-hazing. The plot in Fig. 14 displays the results of various test images de-hazed using mentioned algorithms. From the plot, it is shown that the proposed restoration algorithm is better in visibility enhancement of the hazy images. The last descriptor σ denotes the number of saturated pixels created during visibility restoration.

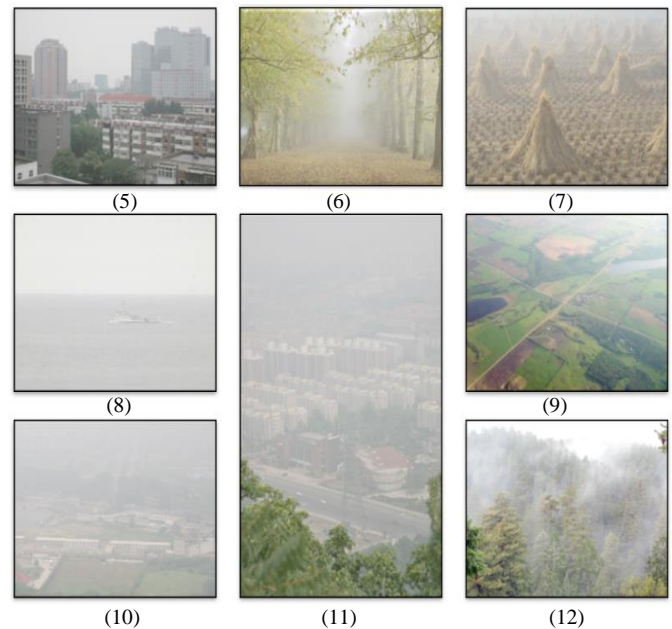


Figure 4: Input Sample Images

The plot of σ for various test images is given in Fig.15. The plot implies that the number of saturated pixels created from the proposed method is lesser than some of the considered algorithms for comparison.

By observing Fig. 5(h) -12(h), one can notice that a good amount of haze has been removed from the input images using the proposed approach. For a 600×400 image the algorithm takes 26.8 seconds for execution using MatlabTM. It is to be noted that out of the 26.8 seconds, 26.1 seconds is taken for smoothing the transmission map. Both subjective and quantitative analysis ascertain the enhanced restoration performance of our proposed method.

7. CONCLUSIONS

This work presents a model based automatic image de-hazing algorithm using the HSV color space. It attempts to map every pixel affected by haze into its nearest true color value in the RGB space; thereby restoring the image. We have proposed an effective way of computing the global atmospheric light. A new method of computing transmission map using saturation and intensity value of hazy image is also presented. The modified haze removal model works well to reliably restore the perceptual visibility of hazy image. Quantitative results using the visible edge segmentation technique demonstrates that the current algorithm is either better or at par with available state of the art methods in removing haze.

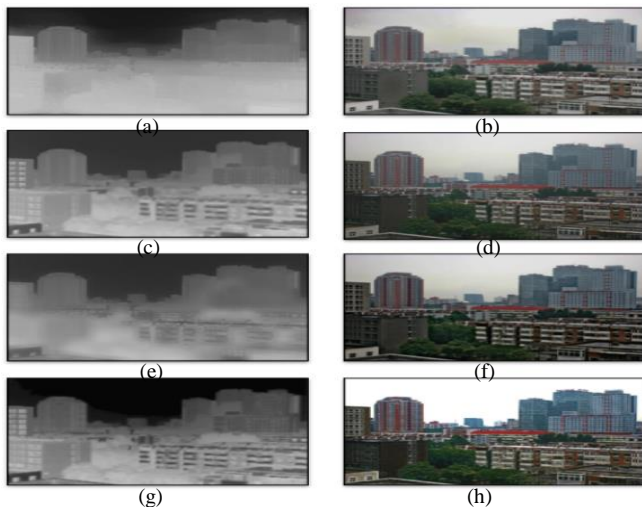


Figure 5: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

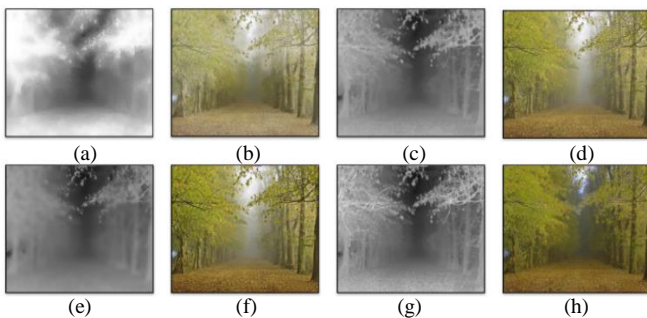


Figure 6: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

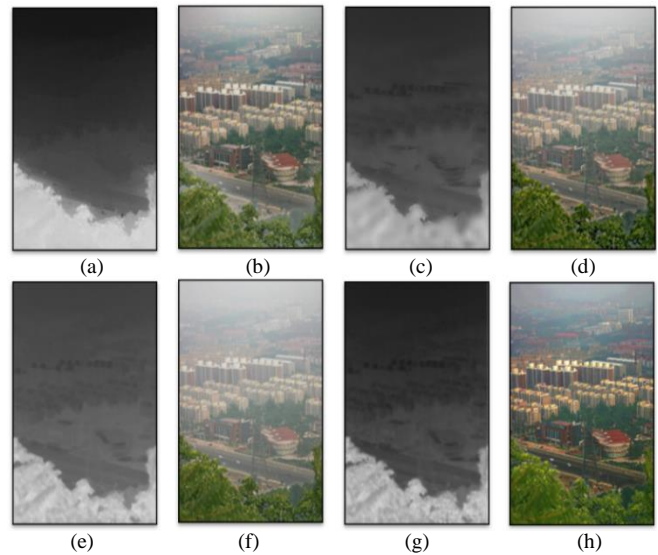


Figure 10: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

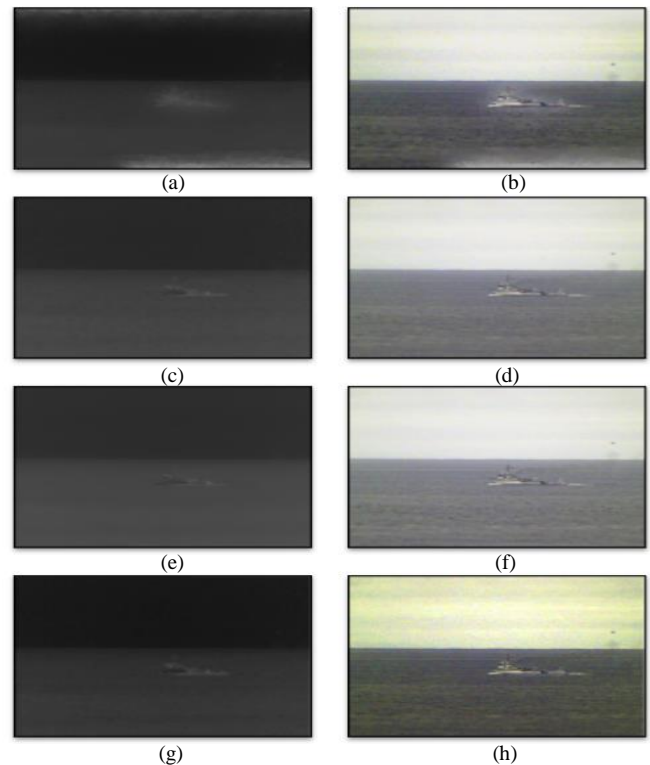


Figure 8: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

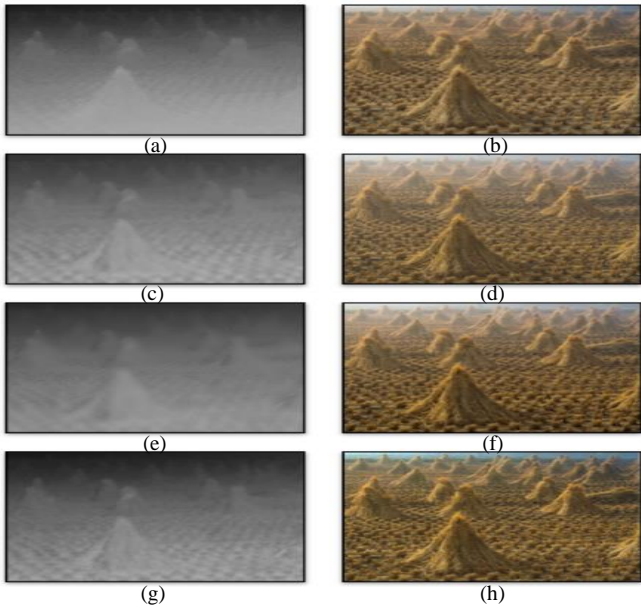


Figure 7: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

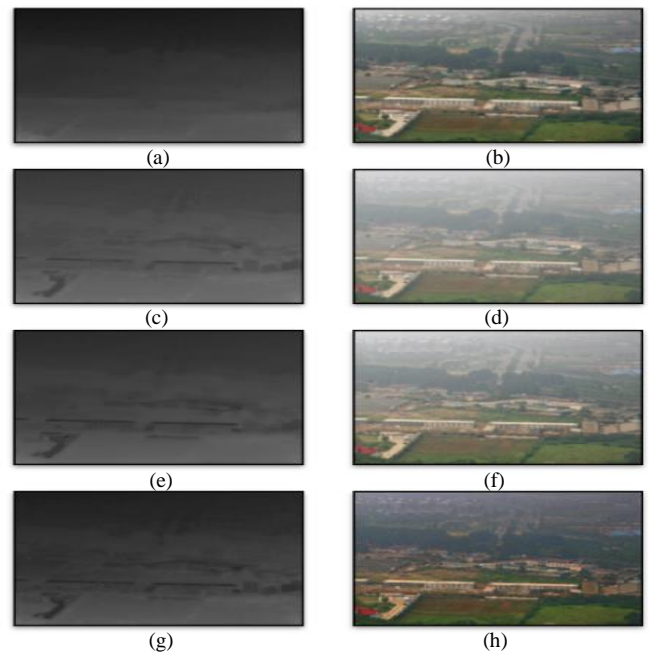


Figure 11: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

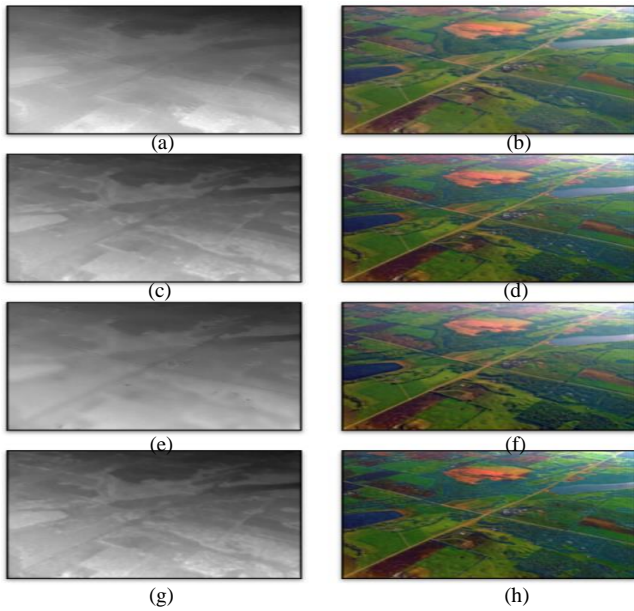


Figure 9: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

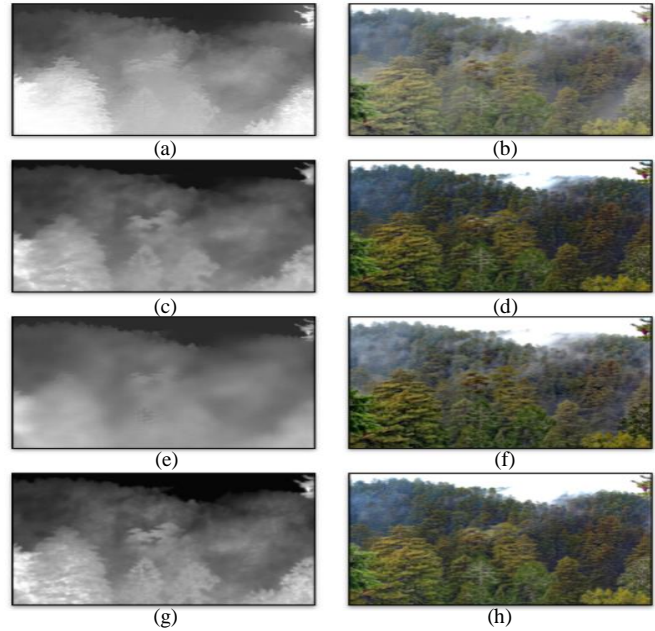


Figure 12: Smoothed transmission map and de-hazed images from the algorithm: He et al. [9] (a & b), Tarel [11] (c & d), Gibson et al. [12] (e & f) and present method (g & h)

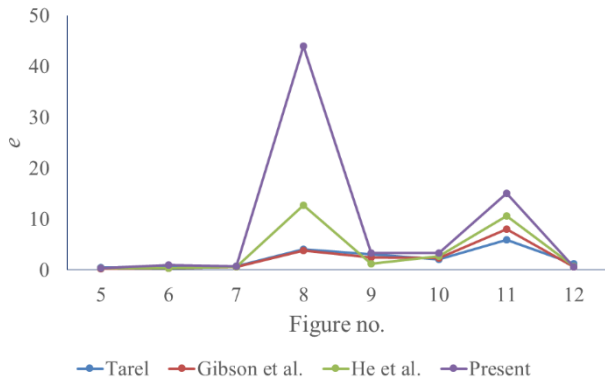


Figure 13. Comparison plot of new visible edge descriptor e for different test images restored using different algorithms.

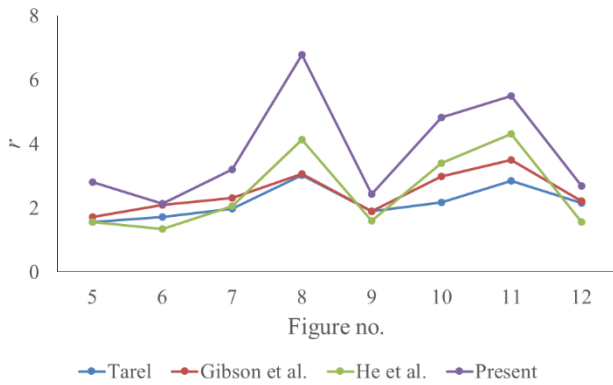


Figure 14. Comparison plot of ratio of visible gradient r for different test images restored using different algorithms.

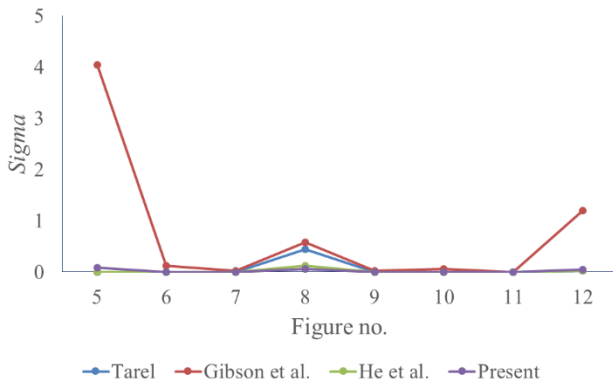


Figure 15. Comparison plot of saturated pixel percentage σ for different test images restored using different algorithms.

8. References

[1] Middleton, W. E. K., "Vision through the Atmosphere", University of Toronto, 1958.

[2] Schechner Y. Y., Narasimhan S. G., Nayar S. K., Instant de-hazing of images using polarization, IEEE conference on Computer Vision and Pattern Recognition, 2001, 1, 325-332.

[3] Narasimhan S. G., Nayar S. K., Chromatic Framework for Vision in Bad Weather, IEEE conference on Computer Vision and Pattern Recognition, 2000, 1, 598-605

[4] Rahman Z., Jobson D. J., Woodell G. A., Retinex Processing for Automatic Image Enhancement, Journal of Electronic Imaging, 2002, 13(1), 568-575.

[5] Narasimhan S. G., Nayar S. K., Interactive Deweathering of an Image Using Physical Models, IEEE workshop color and Photometric Methods in Computer Vision, in Conjunction with IEEE international conference on Computer Vision, 2003.

[6] Tan R. T., Visibility in Bad Weather from a single image, IEEE conference on Computer Vision and Pattern Recognition, 2008, 1-8.

[7] Fattal R., Single Image De-hazing, ACM Transaction on Graphics, 2009, 27(3), 1-9.

[8] Guo F., Cai Z., Xie B., Tang J., Automatic Image Haze Removal Based on Luminance Component, Wireless Communications Networking and Mobile Computing, 2010, 1-4.

[9] He K., Sun J., Tang J., Single Image Haze removal using Dark Channel Prior, IEEE conference on Computer Vision and Pattern Recognition, 2009, 1956-1963.

[10] Zhang Q., Kamata S., Improved Optical Model Based on Region Segmentation for Single Image Haze Removal, Journal of Information and Electronics Engineering, 2012, 2, 62-68.

[11] Tarel J.- P., Hautiere N., Fast Visibility Restoration from a Single Color or Gray Level Image, IEEE conference on Computer Vision, 2009, 2201-2208.

[12] Gibson K. B., Nguyen T. Q., Fast Single Image Fog Removal using the Adaptive Wiener Filter, International Conference on Image Processing, 2013, 714-718.

[13] Lan X., Zhang L., Shen H., Yuan Q., Li H., Single image haze removal considering sensor blur and noise, Journal on Advances in Signal Processing, 2013.

[14] He K., Sun J., Tang X., Guided Image Filtering, Pattern Analysis and Machine Intelligence, 2013, 6, 1397-1409.

[15] Hautiere N., Tarel J.- P., Aubert D., Dumont E., Blind Contrast Enhancement Assessment by Gradient Ratioing at Visible Edges, Journal on Image Analysis & Stereology, 2008, 27(2), 87-95.

Effect of AWGN Parameters Estimation on Accurate Denoising Process

Huda Al-Ghaib

Electrical and Computer Engineering
The University of Alabama in Huntsville
Huntsville, AL 35899
hag0002@eng.uah.edu

Reza Adhami

Electrical and Computer Engineering
The University of Alabama in Huntsville
Huntsville, AL 35899
adhamir@uah.edu

ABSTRACT

The image denoising process attempts to restore a noiseless image from its noisy observation. When the noise is of an unknown source and distribution, it is assumed to have an additive white Gaussian noise (AWGN) distribution. AWGN is characterized by its mean and variance. Denoising digital images of AWGN is a challenging process. An investigation of the relationship between accurate estimation of noise parameters and denoising process is presented in this research. Ant colony optimization (ACO) and region merging algorithms are utilized to estimate noise variance. The denoising process is implemented using wavelet shrinkage operation and the estimated variance. A commonly used metric for measuring the efficiency of denoising algorithms is the peak signal to noise ratio (PSNR). Experimental results based on PSNR measurements showed that an accurate estimation of noise parameters produced better results for the denoising process.

Keywords: Ant colony optimization, region merging, pheromone matrix, k-means clustering, image denoising, wavelet shrinkage process.

i. Introduction

Factors such as moving objects and interference in the transmission channel may occur during image acquisition and/ or transmission [1]-[2]. Applications such as medical imaging, biometrics, pattern recognition, and social media utilize digital images as a main source of information [3]. These applications require high quality noiseless images. To eliminate noise and improve image quality, image denoising is performed on digital images. In most cases, noise originates from an unknown source and location. In this case, the noise is assumed to be an additive white

Gaussian noise (AWGN) of unknown mean and variance. Accurate estimation of noise parameters is a preliminary step in a successful denoising process.

The main goal of the present research is to investigate the relationship between accurate noise estimation and the denoising process. Digital images with AWGN of different densities are considered as input experiments in our research. Two methods to estimate the AWGN noise parameters are presented. Approaches such as ant colony optimization (ACO) and region merging are utilized for this purpose [4], [5]. A wavelet shrinkage operation is performed to denoise the images using the estimated noise parameters. Undecimated wavelet transform (UWT) is applied to analyze the image components prior to the image denoising process.

ii. The Proposed Algorithm

Denoising images with AWGN requires an accurate estimation of noise parameters. These parameters are composed of the mean and variance. In most cases, AWGN is assumed to possess a mean of zero. In our research, two methods are applied to estimate the noise variance: ACO and region merging. In an ACO algorithm, a number of artificial ants are randomly generated on a 2-D graph (image) to locate homogenous regions [4]. In region merging, seed nodes are generated on the image plane and adjacent regions are merged based on a homogeneity test [5]. The denoising process is similar to that of [7].

First, UWT is utilized to decompose an input image into high and low frequency components. A wavelet shrinkage process, explained in section iii, is used in the denoising step. In order to produce a noise free image with reduced artifacts and preserved edges,

denoising is performed on the magnitude of the horizontal and vertical coefficients.

iii. Denoising Algorithm

A wavelet shrinkage denoising operator is stated in [7] as:

$$C(M) = \begin{cases} |M| - T_n, & |M| \geq T_n \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Where M is the magnitude of the detail coefficients. $C(M)$ is a piece wise linear and monotonically non-decreasing function.

T_n is the threshold value. Accurate estimation for T_n is required to have an efficient denoising process. Reference [7] suggested the following mathematical model to compute T_n as:

$$T_n = \sqrt{2\ln(N)}\sigma/\sqrt{N} \quad (2)$$

N is the signal length and σ is the standard deviation of the wavelet coefficients. However, equation (2) cannot be applied to our algorithm for the following reasons:

- The use of a non-orthogonal UWT.
- The shrinkage operation is applied to the magnitudes of the gradient coefficients instead of the wavelet coefficients.

For a white Gaussian noise, the probability distribution function of the magnitude of gradients is characterized by the Rayleigh distribution as:

$$Pr_{||\Delta f||}(m) = \begin{cases} \frac{m}{\eta^2} e^{-\frac{m^2}{2\eta^2}}, & m \geq 0 \\ 0, & m < 0 \end{cases} \quad (3)$$

Where m is a random variable representing the magnitude of the gradient. There is a direct relationship between σ and η . Where σ and η are the standard deviations for Gaussian and Rayleigh distributions respectively. Thus, equation (2) can be rewritten as:

$$T_n = \eta\sqrt{-2\ln(1-p)} \quad (4)$$

p is the probability of noise removal for a particular threshold T_n . $T_n = 3.7169 \eta$ for $p = 0.999$, and $T_n = 4.5\eta$ for $p = 0.99996$ η is the noise variance. Two approaches are utilized to accurately estimate η . The ACO algorithm is explained in detail in section iv, and the region merging algorithm is presented in section v.

iv. Noise Estimation Using Ant Colony Optimization Algorithm

Ants live in colonies and work as a team to search for food, build shelter, and protect their colony. When ants store food, they use the shortest path between the nest and food source. Initially, ants leave the nest using different paths. Each ant leaves a chemical material on the path known as a 'pheromone'. After discovering the shortest path, an ant, X , will use that path numerous times to store food in the nest. Each time X uses the shortest path it will leave a pheromone on the path, and increase its pheromone rate. Other ants will smell the pheromone and follow it. ACO can be applied to digital images to find an optimal solution for a given problem, such as locating homogeneous regions. In this case, the pheromone matrix and heuristic information are used to identify these regions. Initially, a number of artificial ants are generated on a 2-D graph of size $M \times N$, and their paths are traced. When an artificial ant is randomly generated at a position (i, j) it needs to move to (l, m) , where (l, m) is one of the four diagonal nodes. Two parameters determine the next node (l, m) on the ant's path: pheromone and heuristic information. These parameters are computed for each diagonal node. The pheromone matrix is of size $M \times N$ with an initial value of 0.001 in each node. Pheromone value for a node visited by an ant is updated as:

$$w_{l,m}^{(n)} = \begin{cases} (1 - \rho) * w_{l,m}^{(n)} + \rho * \eta_{l,m}, & \text{if a node } (l, m) \text{ is visited} \\ & \text{by artificial ants;} \\ w_{l,m}^{(n)}, & \text{otherwise;} \end{cases} \quad (5)$$

ρ is a constant value used to determine the degree of update at a given node, and $w_{l,m}^{(n)}$ the pheromone value for at the spatial coordinate (l, m) . Using trial and error we found the best value to be $\rho = 0.05$.

The heuristic information η measures the homogeneity of the neighboring nodes. The heuristic information at node (l, m) is illustrated in Fig 1.

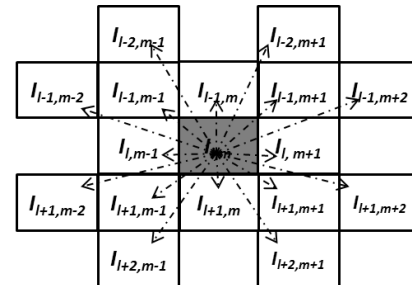


Fig 1. Heuristic information for node (l, m)

Finally, the transition probability is computed for each diagonal node as:

$$p_{(i,j),(l,m)}^{(n)} = \frac{(w_{l,m}^{(n-1)})^\alpha (\eta_{l,m})^\beta}{\sum_{(l,m) \in \Omega_{(i,j)}} (w_{l,m}^{(n-1)})^\alpha (\eta_{l,m})^\beta} \quad (6)$$

$w_{l,m}^{(n-1)}$ is the pheromone value at position (l, m) , $\Omega_{(i,j)}$ is the four diagonal neighboring nodes of (i, j) node. The constants α determines the degree of influence of the pheromone matrix, while β determines the degree of influence of the heuristic matrix. α is a positive value used to increase the influence of the pheromone matrix, while β is a negative value used to decrease the influence of the heuristic matrix. This research utilized values of $\alpha = 1$, and $\beta = -2$. In the final step, the ant moves to the diagonal node (l, m) of the highest transition probability value.

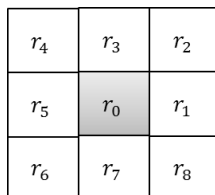
k -means algorithm is applied to classify the pheromone matrix into two vectors composed of most and least frequently visited nodes, i.e., v_1 and v_2 respectively [6]. Only the nodes in vector v_1 are used to estimate the noise variance. Each node in v_1 corresponds to a node (l, m) in the image plane. A patch v_p of size 5×5 centered at node (l, m) is utilized to estimate the local variance. The global variance is the average value of the local variances at different patches.

v. Noise Estimation Using Region-Merging Algorithm

A region-merging algorithm is performed to estimate noise variance. First, an image I is subdivided into $M \times M$ windows. Next, R window is subdivided into Q sub windows as shown in Fig 2. Where r_0, \dots, r_{Q-1} are the sub windows each of size $m \times m$. r_0 is the centered sub window in R . For each sub window in R , the following two conditions must be satisfied:

1. $r_i \cap r_j = \emptyset$, for $i \neq j$.
2. $\cup_i r_i = R$.

We chose $M = 9$, $m = 3$, and $Q = 9$.

Fig 2. Region R of size 9×9

The variance for the pixels within r_0 is σ_0^2 and is considered as the seed sub window in R . Other adjacent sub windows are merged with r_0 if they pass the homogeneity test. The homogeneity test is computed as:

$$h_q = \frac{|\sigma_q^2 - \sigma_0^2|}{\sigma_0^2}, q = 0, 1, \dots, Q - 1 \quad (7)$$

where σ_q^2 is the variance of r_q . Each sub window is assumed to possess a mean of zero, and the variance is computed as:

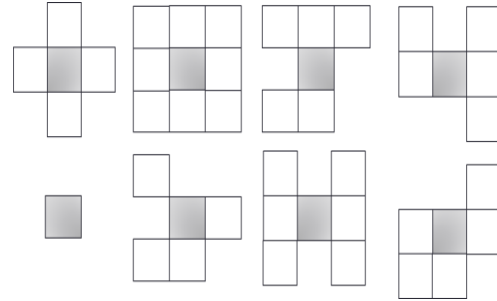
$$\sigma_q^2 = \frac{1}{|c_q|} \sum_{y_m \in c_q} y_m^2 \quad (8)$$

where c_q is the coefficients within r_q . The following condition is applied to test h_q :

$$m_q = \begin{cases} 1, & \text{if } h_q < t \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where $t = 0.2$. If $m_q = 1$ for a given r_q , then r_q is merged with r_0 , otherwise it is discarded. The homogeneity test is performed on every sub window in R . As a result, final R is of an arbitrary shape. When $Q = 9$, there is $2^{Q-1} = 256$ different configurations for R . Fig 3 shows some configurations for R .

The local noise variance is estimated using R with an arbitrary shape (configuration). The global variance is the average value for the local variances.

Fig 3. Local windows with arbitrary size and shape, r_0 is the gray-scale region

vi. Error Rate for σ estimation

A set of images is used as experimental input to test the noise variance estimation algorithms. These images are: Lena, cameraman, barbara, kodim05, kodim06, kodim07, kodim08, kodim21, and kodim24. AWGN with different standard deviation values are added to the input images, i.e., $\sigma = 80.6, 71.4, 57.51$ and 25.5 respectively. This produces images of high noise density. Table I displays the averaged estimation normalized values

for the noise variance using ACO and NRM (Noise estimation using Region Merging).

σ^2 (Added)	0.1	0.078	0.05	0.04	0.01
σ^2 (Estim.)					
ACO	0.082	0.068	0.059	0.048	0.028
NRM	0.074	0.062	0.045	0.038	0.014

Table I. Noise variance estimation

Table I shows that for high noise densities, i.e., 0.1 and 0.078, ACO algorithm provided better estimation. However, for low noise densities, i.e., 0.04 and 0.01, NRM provided better estimation for noise variance compared with ACO. For noise density= 0.05, the two algorithms provided similar results.

vii. Experimental Results

The noisy images produced in section vi are fed to a wavelet shrinkage algorithm to denoise them. The denoising process is performed twice using the resultant estimated variances from ACO and NRM algorithms respectively. Peak signal to noise ratio (PSNR) is computed for the noisy and denoised images. Table II shows the results for the denoising step. The average value of the estimated variances for the input images is shown in Table II.

σ^2 (Added)	0.1	0.078	0.05	0.04	0.01
PSNR-noisy	14.156	14.907	16.534	17.424	23.183
PSNR-denoised (ACO)	22.186	23.087	24.071	24.447	27.526
PSNR-denoised (NRM)	21.900	22.962	24.127	24.452	27.974

Table II. PSNR for noisy and denoised images

From data in Table II, it is obvious that for high noise densities denoising using ACO estimated variances provided higher values for PSNR, while for low noise densities NRM estimated variances provided better results. This is due to the fact that ACO algorithm produced more accurate estimation for images with high noise densities. Meanwhile, NRM was superior for low noise densities as can be seen in Table I.

viii. Conclusion

In this research optimization of AWGN parameters for image noise estimation and denoising process is presented. The noise estimation and elimination processes are implemented on digital

images containing AWGN of high noise density. Ant colony optimization and region merging algorithms are applied to estimate noise variance and a wavelet shrinkage operation is utilized to denoise the images. The peak signal to noise ratio (PSNR) was used to compare the performance. Experimental results illustrated in Table II shows that accurate noise estimation yielded a more effective denoising process. PSNR is increased when the noise variance is accurately estimated for all the input test images using different noise densities.



Fig 4. Denoised image with PSNR = 32.4286

ix. References

- [1] "Mitigating motion artifact in FDK based 3D Cone-beam Brain Imaging System using markers," Ujjal K. Bhowmik, M. ZafarIqbal, and Reza R. Adhami, Central European Journal of Engineering, DOI: 10.2478/s13531-012-0011-7.
- [2] R. C. Gonzalez, R. E. Woods, "Digital Image Processing", 3rd edition, 2008.
- [3] Al-Ghaib, H.; Adhami, R., "An E-learning interactive course for teaching digital image processing at the undergraduate level in engineering," *Interactive Collaborative Learning (ICL), 2012 15th International Conference on*, vol., no., pp.1,5, 26-28 Sept. 2012.
- [4] Jing Tian; Li Chen, "Image Noise Estimation Using A Variation-Adaptive Evolutionary Approach," *Signal Processing Letters, IEEE*, vol.19, no.7, pp.395,398, July 2012.
- [5] Il Kyu; Yoo Shin Kim, "Wavelet-based denoising with nearly arbitrarily shaped windows," *Signal Processing Letters, IEEE*, vol.11, no.12, pp.937,940, Dec. 2004.
- [6] R. O. Duda, P. E. Hart, D. G. Stork, "Pattern Classification", 2nd edition, 2000.
- [7] J. Fan and A. Laine, "Contrast enhancement by multiscale and nonlinear operators," *Wavelets in Medicine and Biology*. Boca Raton, FL: CRC, Mar. 1996, pp. 163-192.
- [8] Mencattini, A.; Salmeri, M.; Lojacono, R.; Frigerio, M.; Caselli, F., "Mammographic Images Enhancement and Denoising for Breast Cancer Detection Using Dyadic Wavelet Processing," *Instrumentation and Measurement, IEEE Transactions on*, vol.57, no.7, pp.1422,1430, July 2008.

DCT-BASED IMAGE QUALITY ASSESSMENT FOR MOBILE SYSTEM

Jeong Sung Park and Tokunbo Ogunfunmi

Department of Electrical Engineering
Santa Clara University
Santa Clara, CA 95053, USA

Email: jeoongsung@gmail.com and togunfunmi@scu.edu

ABSTRACT

In this paper, we do further research on the DCT-based image quality approach proposed in our previous paper [7]. Our objective is to find a new image quality metric that run on the fly at video encoder and decoder in mobile systems. Most of dominant image quality metrics such as SSIM use intensity, mean, variance, and covariance on the pixel domain which take too much hardware and complexity. Instead, we propose to measure just frequency difference between original image and distorted image. It takes low complexity enough to implement on hardware. By using a built-in DCT block in image and audio standards such as H.264 and HEVC, much hardware for computing frequency components can be saved. In this paper, we propose a performance-improved metric than FSM (Frequency Similarity Method) proposed in [7]. As a result of simulation, our proposed metric performs by 95 percent as SSIM does. Even though 95 percent is not better than SSIM, it is enough to make video system more adaptive and rate-controllable based on error measurement.

Index Terms— Objective image quality assessment, FMSE, FSM

1. INTRODUCTION

Image quality is a characteristic of an image that measures the perceived image degradation. It plays an important role in various image processing application. Goal of image quality assessment is to supply quality metrics that can predict perceived image quality automatically. There are two types of image quality assessment: subjective quality assessment and objective quality assessment ([1], [2], [3]).

Subjective image quality is concerned with how image is perceived by a viewer and gives his or her opinion on a particular image. The mean opinion score (MOS) has been used for subjective quality assessment. Objective image quality assessment is a mathematical model that approximates results of subjective quality assessment. Goal of objective evaluation is to develop quantitative measure that can predict perceived image quality. MSE (Mean Square Error) and PSNR (Peak

Signal to Noise Ratio) ([12], [13]) are the most used methods of objective image quality assessment for image quality assessment. It measures pixel-to-pixel error between a reference image and a distorted image.

Alternatively Wang et al. ([4]) proposed the Structural Similarity (SSIM) index. This method extracts the structural information of the image and has been proved to be a new representative metric that reflects HVS. However, one may think of situations in which the information provided by this index does not match a subjective quality judgement. It is due to the bias each method has towards the image statistic it is using to measure. Some other quality assessment methods based on different features may give more accurate information of the global quality. [6] reported a drawback of SSIM and presented the Quality Index based on Local Variance (QILV) which is a new method based on the distribution of the local variance in the images with the aim to better handle the non-stationarity of the images to be compared. [5] proposes to add frequency structural comparison onto SSIM. But, frequency information is used redundantly and calculation is very complicated. [11] develops a general-purpose no-reference approaches to image quality assessment based on a DCT statistics.

Our objective is to find a new image quality metric that run on the fly at video encoder and decoder in mobile systems. If it is possible to measure BER in the receiver without taking CPU and requiring much hardware resource, mobile systems can become more intelligent and adaptive. Most of dominant image quality metrics such as SSIM use intensity, mean, variance, and covariance on the pixel domain which take too much hardware and complexity.

In our previous paper [7], we proposed a new image quality assessment which was named as FSM (Frequency Similarity Method). FSM measures just frequency difference between a distorted image and a reference image by using Discrete Cosine Transform (DCT). It takes low complexity enough to implement on hardware. By using a built-in DCT block in image and audio standards such as H.264 and HEVC, much hardware for computing frequency components can be saved. Experimental result showed FSM achieved 90 percent

performance of SSIM in [7] which was higher than 86 percent of PSNR [5].

In this paper, we propose FMSE (Frequency Mean Square Error) as a new definition of FSM to get more precise image quality metric. Based on experimental results with standard image database ([9]), FMSE achieves 95 percent performance of SSIM. Besides, FMSE performs better than SSIM especially at white-noised images. We also explore various transform sizes such as 16×16 , 8×8 and 4×4 to get better performance.

This paper is organized as follows. Section II presents a related background about image quality metric. Section III describes our proposed method. In Section IV, experimental results are provided and we conclude in Section V.

2. BACKGROUND

In this section, we present a brief overview of image quality metric. MSE (Mean Square Error) and PSNR (Peak Signal to Noise Ratio) ([12], [13]) measure pixel-to-pixel error between a reference image and a distorted image as denoted in Equations (1) and (2).

$$MSE = \frac{1}{MN} \sum_{j=1}^M \sum_{j=1}^N (x_{ij} - y_{ij})^2 \quad (1)$$

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \quad (2)$$

where L is a maximum level of intensity. As presented in [4], equations of SSIM are as follows.

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (3)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (4)$$

$$s(x, y) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (5)$$

Equation (3), (4), and (5) represent contrast comparison, luminance comparison, and structure similarity comparison, respectively. In Equation (5), structural similarity comparison is given by using covariance of x and y and both variance of x and variance of y . At last, equation (6) includes all those comparisons.

$$s(x, y) = \frac{(\mu_x\mu_y + C_1)(\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (6)$$

[6] reported some drawbacks of SSIM. Figure 1 which is obtained from [6] shows 4 different Lena images Figure 1-(b) to Figure 1-(e) that have the same SSIM=0.50 with reference image Figure 1-(a). All 4 images can not be perceived to human visual system with the same feeling and level. For example,

(b) looks definitely better than Figure 1-(c). Figure 1-(e) can not be identified as Lena image without any information. As shown in Figure 1, an important drawback of SSIM is a bias towards some features of the image. SSIM is too sensitive to white noise and speckle noise and too generous to blurring.

In our previous paper [7], our proposed method which is named as FSM (Frequency Similarity Method) estimates similarity between the frequency map of a reference image and that of a distorted image. To transform pixel data to frequency domain, it uses 8×8 block based DCT for simple calculation.

$$FSM_i = \frac{\min(X_i, Y_i) + C}{\max(X_i, Y_i) + C} \quad (7)$$

where X and Y are transformed results of original image and distorted image, respectively. i is a position index on a new transformed image which has the same size as the original image. C is a small constant used to avoid instability when the denominator might approach zero. Equation (7) indicates relative difference between frequency components of the original image and the distorted image regardless of which one is greater. Mean value of FSM_i over all positions is the final metric for image quality assessment as shown in Equation (8).

$$FSM = \frac{1}{N} \sum_{i=1}^N \frac{\min(X_i, Y_i + C)}{\max(X_i, Y_i + C)} \quad (8)$$

where $0 \leq FSM \leq 1$. N is the number of pixels. Compared with Equation (6), Equation (8) is much simpler than other equations.

3. IMAGE QUALITY ASSESSMENT BASED ON FREQUENCY SIMILARITY

In [7], FSM achieves 90 percent performance of SSIM. That performance is OK to detect trend of low errors or high error by using minimum hardware size. But, our new target is to make system performance higher than 90 percent of SSIM. To obtain higher performance, difference of each frequency component between original image and distorted image needs to be measured more precisely. One of popular and precise methods is MSE. So, we obtain frequency components of the original image and the distorted image by using DCT and apply them into MSE as follows.

$$FMSE = MSE(X, Y) = \frac{1}{MN} \sum_{j=1}^M \sum_{j=1}^N (X_{ij} - Y_{ij})^2 \quad (9)$$

where all variables and index have the same meaning as Equation (8). FMSE simply indicates MSE on frequency domain. Compared with Equation (8), Equation (9) can measure more statistical frequency difference and does not have dependency on C . Of course, complexity of Equation (9) is low enough to implement simply on hardware compared with Equation (6).

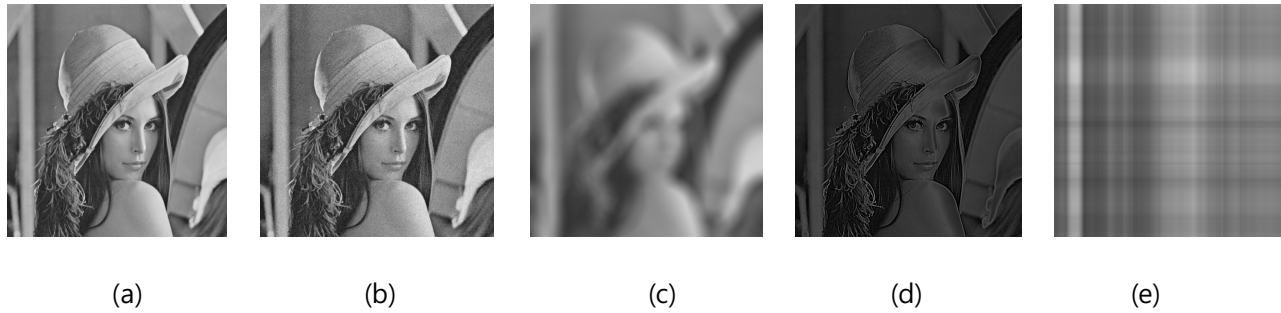


Fig. 1. (a) Original Image. Other images have the same SSIM=0.582 (b) white noise added, (c) blur distortion (d) high-boosted (e) singular value decomposition (most significant eigenimage). This figures are referred from [6]

In the same as FSM, FMSE can use a built-in DCT block for data compression. If the original DCT block in video system is used for FSM, additional hardware resource is not needed to obtain frequency components. Only a few more additional operation units for multiplication, addition, and division are required for calculating Equation (9).

4. EXPERIMENTAL RESULTS

4.1. Similarity and difference between FMSE and SSIM

Figure 2 shows relation between FMSE and SSIM. Each quality of image could be generated by two ways. One way is adding speckling noise. The other way is blurring. There are two plots that correspond to those two ways. As shown in both figures, FMSE has very high correlation with SSIM even though its complexity is lower than SSIM.

However, the curve in Figure 2-(a) is sharper than that of Figure 2-(b). For example, in case of points corresponding to SSIM 0.6 on both plots, FMSE value of the blurred Lena image is 580 which indicates SSIM is too generous to blurring. But, FMSE value of the speckle-noised Lena image is 120 which indicates SSIM is too sensitive to white noise and speckle noise. In the other hand, FMSE shows more balanced and less biased results against the drawbacks of SSIM. As denoted in [7], frequency-based image quality metrics are robust against the drawbacks of SSIM. In Figure 1, FMSE value of each image is 125, 950, 890, and 1240, respectively.

4.2. LIVE database

The LIVE Image Quality Assess Database [9], together with the subjective score for each image was used to validate the performance of the proposed algorithm. In order to provide quantitative measures on the performance of the objective quality assessment models, we follow the performance evaluation procedures provided by the video quality experts group (VQEG) Phase II FR-TV test [10]. From [10], the logistic functions are applied in fitting procedure to provide a nonlin-

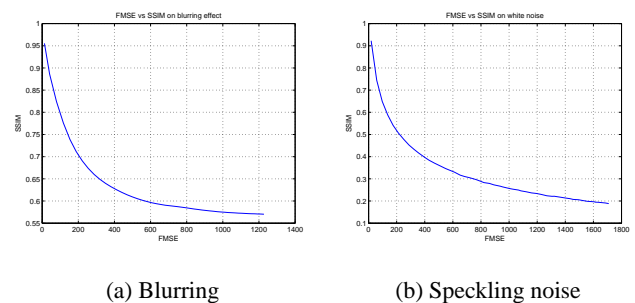


Fig. 2. Correlation between FSM and SSIM

ear mapping between the objective/subjective scores as shown in Figures 1 and 3. Then, Metric1 (The Pearson linear correlation coefficient) and Metric2 (Spearman rank order correlation coefficient) are used for comparison. FSM in Figures 3 to 4 and Tables 1 to 4 indicates FMSE. Tables 1, 2, and 3 show the quantitative results.

There are 5 groups in LIVE database images : JPEG, JPEG2000, Fast Fading, White Noise, and Gaussian Blur. We divide those 5 groups into two groups which are the first three groups and the last two groups. In case of the first group, logistic regression curves between DMOS and SSIM/FMSE could be obtained easily as shown in Figure 3. However, we failed to obtain those of the second group. There were many outliers among DMOS values in LIVE database images. So, we changed comparison target of the second group (White Noise and Gaussian Blur) from DMOS to standard deviation values which are also included in LIVE database images. As a logistic regression result, we could obtain better curve than DMOS as shown in 4. When we remove outliers from all images in the second group, the logistic regression curve of DMOS is very similar to that of standard deviation. To compare more sample images, we select standard deviation values rather than DMOS values for the second group. Tables 2 and 3 indicate SSIM performs higher by 5 percent than FMSE at all groups other than the White Noise group. Exceptionally,

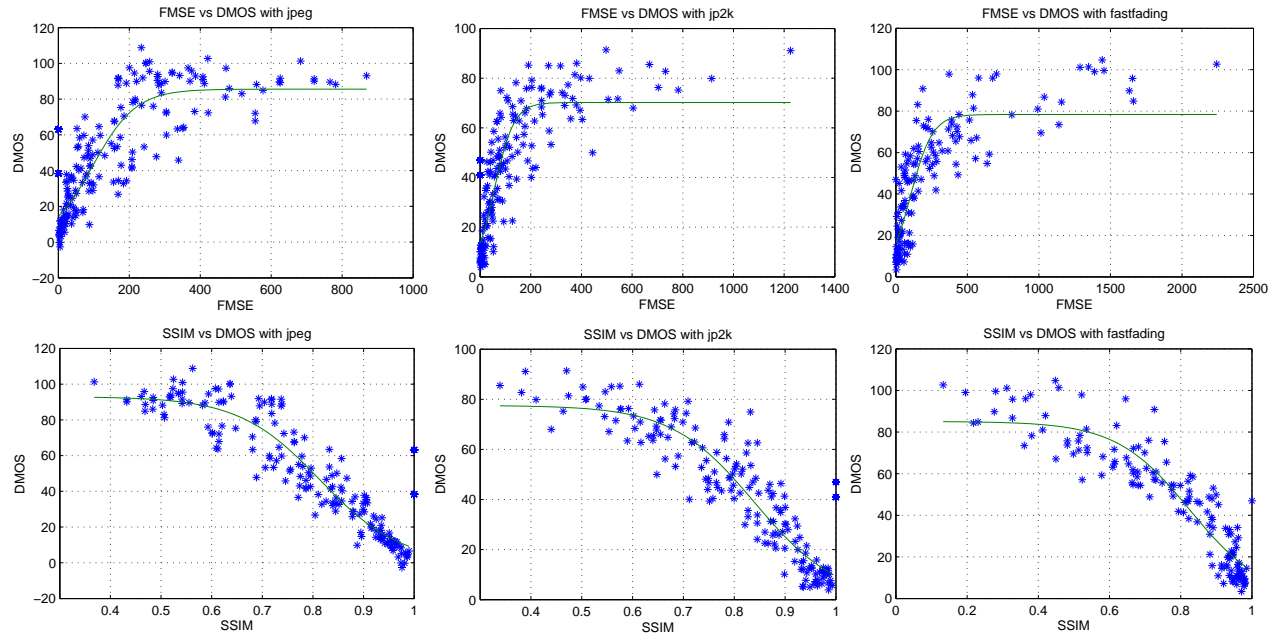


Fig. 3. Logistic regression curves with JPEG, JPEG2000, and Fast Fading images

MODEL	JPEG2000					JPEG					Fast Fading				
	SROCC	CC	MAE	RMS	OR	SROCC	CC	MAE	RMS	OR	SROCC	CC	MAE	RMS	OR
SSIM	0.9350	0.9345	7.1170	9.0007	0.0473	0.9436	0.9475	7.9596	10.1900	0.0686	0.9403	0.9304	8.6305	10.4693	0.0414
FMSE	0.8956	0.8831	9.2754	11.8557	0.0533	0.8814	0.8828	11.9661	14.9772	0.0629	0.8949	0.8703	11.6506	14.0427	0.0414

Table 1. Performance comparison of SSIM and FMSE on JPEG, JPEG2000, Fast Fading images

FMSE performs higher by 4 percent than SSIM at the White Noise group.

As experimental results on all LIVE database images in Table 3, FMSE achieves 95 percent performance as SSIM does. Even though FMSE does not outperform SSIM, it has more practical advantages: lower complexity, less hardware resource, and easy adaptation to existing video systems compared with SSIM as mentioned in section 3. Table 4 compares performance of FMSE with transform sizes of 16×16 , 8×8 and 4×4 . As the block size of DCT gets larger, performance of FMSE gets a bit higher. This is because a larger size of DCT block includes more frequency components than a smaller one. But, performance difference is very small.

5. CONCLUSIONS

We presented an improved DCT-based metric for image quality assessment named as FMSE. FMSE estimates similarity of frequency between a distorted image and a reference image using DCT. Different from SSIM and PSNR, it does not use any data on the pixel domain. Instead, it simply uses only frequency components. Experimental results show FMSE

achieves 95 percent performance as SSIM does. That is, it still has very high correlation with subject scores (DMOS). Besides, the computational complexity of FMSE is simpler than SSIM. Since it uses a fixed number of coefficients for matrix multiplication, it can be easily implemented on hardware. FMSE can use the DCT block which is already built in video system. If the original DCT block in video system is used for FMSE, simple additional hardware resource is needed to measure image quality. Mobile system (video encoder or decoder) can become more intelligent and adaptive by using FMSE for image quality assessment on the fly.

6. REFERENCES

- [1] K.R. Rao and H. R. Wu, "Digital Video Image Quality and Perceptual Coding," CRC Press, 2006.
- [2] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *Broadcasting, IEEE Transactions on*, vol.54, no.3, pp.660-668, Sept. 2008.

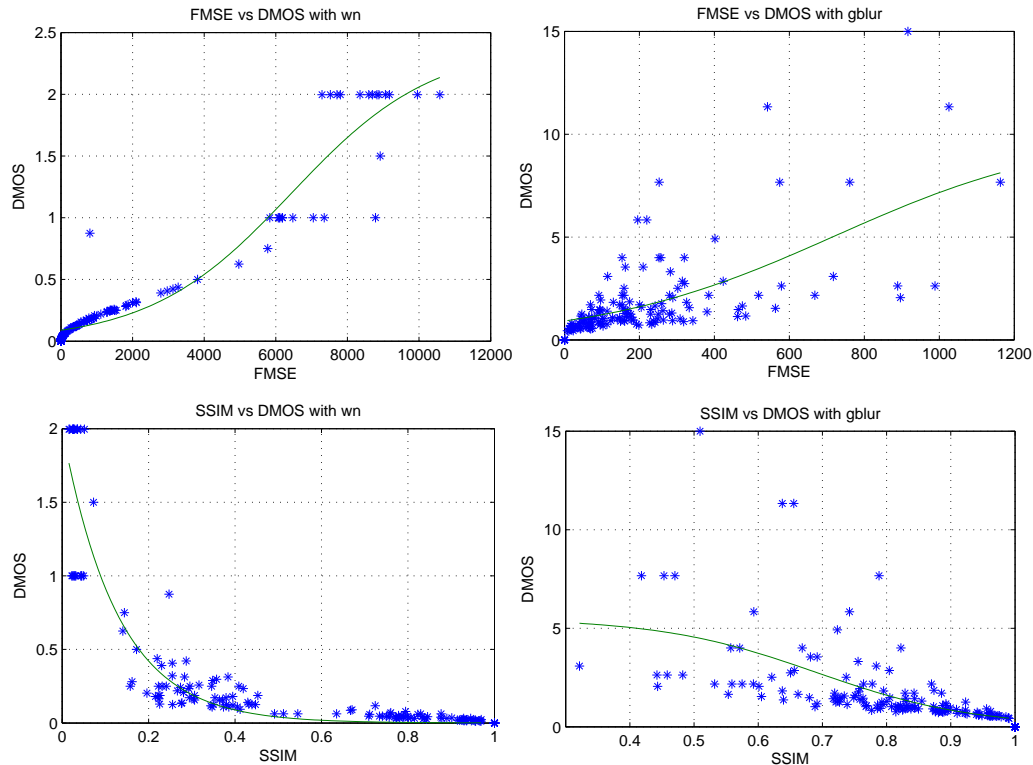


Fig. 4. Logistic regression curves with white noise and gaussian blur images

MODEL	White Noise					Gaussian Blur				
	SROCC	CC	MAE	RMS	OR	SROCC	CC	MAE	RMS	OR
SSIM	0.9543	0.9301	0.1305	0.2146	0.0828	0.8507	0.5808	0.9024	1.7284	0.0345
FMSE	0.9961	0.9646	0.0891	0.1529	0.0621	0.7251	0.6405	0.9499	1.6307	0.0690

Table 2. Performance comparison of SSIM and FMSE on white noise and gaussian blur images

- [3] Z. Wang, A. C. Bovik, "Modern Image Quality Assessment," *New York: Morgan and Claypool Publishing Company*, 2006.
- [4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600-612, Apr. 2004.
- [5] D. Lv, D. Bi, and Y. Wang, "Image Quality Assessment Based on DCT and Structural Similarity," *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, vol., no., pp. 1-4, 23-25 Sept. 2010.
- [6] S. Aja-Fernandez, R. San Jose Estepar, C. Alberola-Lopez, and C.F. Westin, "Image quality assessment based on local variance," *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, vol., no., pp. 4815-4818, Aug. 30-Sept. 3 2006.
- [7] J.S. Park and T. Ogunfunmi, "A New Approach for Image Quality Assessment: Frequency Similarity Method (FSM)," *Proceedings of the IEEE International Conference on Industrial Electronics (ICIEA)*, Singapore, July 2012.
- [8] J.S. Park and T. Ogunfunmi, "Image quality assessment using frequency similarity," *Provisional Patent Filing*, USA, June 2012.
- [9] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, LIVE Image Quality Assessment Database Release 2 [Online]. Available: <http://live.ece.utexas.edu/research/quality> 2006
- [10] VQEG. Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video

MODEL	JPEG, JPEG2000, Fast Fading					White Noise and Gaussian Blur				
	SROCC	CC	MAE	RMS	OR	SROCC	CC	MAE	RMS	OR
SSIM	0.9397	0.9375	7.9024	9.8866	0.0524	0.9025	0.7555	0.5165	9.8866	0.0586
FMSE	0.8907	0.8787	10.9640	13.6252	0.0525	0.8606	0.8026	0.5195	13.6252	0.0655

Table 3. Average values of SROCC, CC, MAE, RMS, OR

All images					
MODEL	SROCC	CC	MAE	RMS	OR
4x4	0.8746	0.8395	5.7701	7.2868	0.0579
8x8	0.8748	0.8398	5.7655	7.2809	0.0589
16x16	0.8756	0.8406	5.7418	7.2585	0.0590

Table 4. Performance of FMSE with 4x4, 8x8, and 16x16

Quality Assessment. Phase II (FR-TV2)(2003, 9). Available: <http://www.vqeg.org/>

- [11] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT Statistics-Based Blind Image Quality Index," *IEEE Signal Processing Letters*, Vol. 17, No. 6., pp. 583-586, June 2010.
- [12] H. Tang and L. Cahill, "A new criterion for the evaluation of image restoration quality," *TENCON '92. 'Technology Enabling Tomorrow : Computers, Communications and Automation towards the 21st Century.'* 1992 *IEEE Region 10 International Conference.*, vol.2, pp.573-577, 11-13 Nov. 1992.
- [13] A. Eskicioglu and P. Fisher, "Image quality measures and their performance," *Communications, IEEE Transactions on*, vol. 43, no. 12, pp. 2959-2965, Dec. 1995.

χ SET: Image Coder based on Contrast Band-Pass Filtering

Jaime Moreno^{†‡}, Oswaldo Morales[†], and Ricardo Tejada[†]

[†]Superior School of Mechanical and Electrical Engineers, National Polytechnic Institute of Mexico, IPN Avenue, Lindavista, Mexico City, 07738, Mexico.

[‡]Signal, Image and Communications Department, University of Poitiers, Poitiers, 30179, France.
e-mail:jmorenoe@ipn.mx

Abstract—Noise is fatal to image compression performance because it can be both annoying for the observer and consumes excessive bandwidth when the imagery is transmitted [1], [2]. Some noise, in addition to some numerical redundancy, is removed during the quantization process, but in some circumstances the removed information is easily perceived by the observer, leading to annoying visual artifacts. Perceptual quantization reduces unperceivable details and thus improve both visual impression and transmission properties. In this work, we apply perceptual criteria in order to define a perceptual forward and inverse quantizer. It is based on the CBPF, a low-level computational model that reproduces color perception in the Human Visual System. Our approach consists in performing a local quantization of wavelet transform coefficients using some of the human visual system behavior properties. It is performed applying a local weight for every coefficient. The CBPF allows to recover these weights from the quantized data, which avoids the storing and transmission of these weights. We apply this perceptual quantizer to the Hi-SET coder. The comparison between JPEG2000 coder and the combination of Hi-SET with the proposed perceptual quantizer (χ SET) is shown. The latter produces images with lower PSNR than the former, but they have the same or even better visual quality when measured with well-known image quality metrics such as MSSIM, UQI or VIF, for instance. Hence, χ SET obtain more compressed (i.e. lower bit-rate) images at the same perceptual image quality than JPEG2000. Keywords: Hu-

man Visual System, Contrast Sensitivity Function, Perceived Images, Wavelet Transform, Peak Signal-to-Noise Ratio, No-Reference Image Quality Assessment, JPEG2000.

1. Introduction

Digital image compression has been a research topic for many years and a number of image compression standards has been created for different applications. The JPEG2000 is intended to provide rate-distortion and subjective image quality performance superior to existing standards, as well as to supply functionality [3]. However, JPEG2000 does not provide the most relevant characteristics of the human visual system, since for removing information in order to

compress the image mainly information theory criteria are applied. This information removal introduces artifacts to the image that are visible at high compression rates, because of many pixels with high perceptual significance have been discarded. Hence, it is necessary an advanced model that removes information according to perceptual criteria, preserving the pixels with high perceptual relevance regardless of the numerical information. The Chromatic Induction Wavelet Model presents some perceptual concepts that can be suitable for it. Both CBPF and JPEG2000 use wavelet transform. CBPF uses it in order to generate an approximation to how every pixel is perceived from a certain distance taking into account the value of its neighboring pixels. By contrast, JPEG2000 applies a perceptual criteria for all coefficients in a certain spatial frequency independently of the values of its surrounding ones. In other words, JPEG2000 performs a global transformation of wavelet coefficients, while CBPF performs a local one. CBPF attenuates the details that the human visual system is not able to perceive, enhances those that are perceptually relevant and produces an approximation of the image that the brain visual cortex perceives. At long distances, the lack of information does not produce the well-known compression artifacts, rather it is presented as a softened version, where the details with high perceptual value remain (for example, some edges).

2. Chromatic Induction Wavelet Model: Brief description.

The Chromatic Induction Wavelet Model (CBPF) [4] is a low-level perceptual model of the HVS. It estimates the image perceived by an observer at a distance d just by modeling the perceptual chromatic induction processes of the HVS. That is, given an image \mathcal{I} and an observation distance d , CBPF obtains an estimation of the perceptual image \mathcal{I}_p that the observer perceives when observing \mathcal{I} at distance d . CBPF is based on just three important stimulus properties: spatial frequency, spatial orientation and surround contrast. This three properties allow to unify the chromatic assimilation and contrast phenomena, as well as some other perceptual processes such as saliency perceptual processes.

The perceptual image \mathcal{I}_p is recovered by weighting these $\omega_{s,o}$ wavelet coefficients using the *extended Contrast Sen-*

sitivity Function (e-CSF). The e-CSF is an extension of the psychophysical CSF [5] considering spatial surround information (denoted by r), visual frequency (denoted by ν , which is related to spatial frequency by observation distance) and observation distance (d). Perceptual image \mathcal{I}_ρ can be obtained by

$$\mathcal{I}_\rho = \sum_{s=1}^n \sum_{o=v,h,dgl} \alpha(\nu, r) \omega_{s,o} + c_n, \quad (1)$$

where $\alpha(\nu, r)$ is the e-CSF weighting function that tries to reproduce some perceptual properties of the HVS. The term $\alpha(\nu, r) \omega_{s,o}$ is considered the *perceptual wavelet coefficients* of image \mathcal{I} when observed at distance d and is written as:

3. JPEG2000 Global Visual Frequency Weighting

In JPEG2000, only one set of weights is chosen and applied to wavelet coefficients according to a particular viewing condition (100, 200 or 400 dpi's) with fixed visual weighting [3, Annex J.8]. This viewing condition may be truncated depending on the stages of embedding, in other words at low bit rates, the quality of the compressed image is poor and the detailed features of the image are not available since at a relatively large distance the low frequencies are perceptually more important.

The table 1 specifies a set of weights which was designed for the luminance component based on the CSF value at the mid-frequency of each spatial frequency. The viewing distance is supposed to be 4000 pixels, corresponding to 10 inches for 400 dpi print or display. The weight for LL is not included in the table, because it is always 1. Levels 1, 2, ..., 5 denote the spatial frequency levels in low to high frequency order with three spatial orientations, *horizontal*, *vertical* and *diagonal*.

Table 1

RECOMMENDED JPEG2000 FREQUENCY (s) WEIGHTING FOR 400 DPI'S
($s = 1$ IS THE LOWEST FREQUENCY WAVELET PLANE).

s	<i>horizontal</i>	<i>vertical</i>	<i>diagonal</i>
1	1	1	1
2	1	1	0.731 668
3	0.564 344	0.564 344	0.285 968
4	0.179 609	0.179 609	0.043 903
5	0.014 774	0.014 774	0.000 573

4. Perceptual Forward Quantization

4.1 Forward

Quantization is the only cause that introduces distortion into a compression process. Since each transform sample at the perceptual image \mathcal{I}_ρ (1) is mapped independently to a

corresponding step size either Δ_s or Δ_n , thus \mathcal{I}_ρ is associated with a specific interval on the real line. Then, the perceptually quantized coefficients \mathcal{Q} , from a known viewing distance d , are calculated as follows:

$$\mathcal{Q} = \sum_{s=1}^n \sum_{o=v,h,dgl} \text{sign}(\omega_{s,o}) \left[\frac{|\alpha(\nu, r) \cdot \omega_{s,o}|}{\Delta_s} \right] + \left[\frac{c_n}{\Delta_n} \right] \quad (2)$$

Unlike the classical techniques of Visual Frequency Weighting (VFW) on JPEG2000, which apply one CSF weight per sub-band [3, Annex J.8], Perceptual Quantization using CBPF (ρ SQ) applies one CSF weight per coefficient over all wavelet planes $\omega_{s,o}$. In this section we only explain Forward Perceptual Quantization using CBPF (F- ρ SQ). Thus, (2) introduces the perceptual criteria of the Perceptual Images (1) to each quantized coefficient of the Dead-zone Scalar Quantizer[3, Annex J.8]. A normalized quantization step size $\Delta = 1/128$ is used, namely the range between the minimal and maximal values at \mathcal{I}_ρ is divided into 128 intervals. Finally, the perceptually quantized coefficients are entropy coded, before forming the output code stream or bitstream.

4.2 Inverse

The proposed Perceptual Quantization is a generalized method, which can be applied to wavelet-transform-based image compression algorithms such as EZW, SPIHT, SPECK or JPEG2000. In this work, we introduce both forward (F- ρ SQ) and inverse perceptual quantization (I- ρ SQ) into the *Hi-SET* coder [6], [7], [8]. An advantage of introducing ρ SQ is to maintain the embedded features not only of *Hi-SET* algorithm but also of any wavelet-based image coder. Thus, we call Perceptual Quantization + *Hi-SET* = *PHi-SET* or χ SET.

Both JPEG2000 and χ SET choose their VFWs according to a final viewing condition. When JPEG2000 modifies the quantization step size with a certain visual weight, it needs to explicitly specify the quantizer, which is not very suitable for embedded coding. By contrast, χ SET neither needs to store the visual weights nor to necessarily specify a quantizer in order to keep its embedded coding properties.

The main challenge underlies in to recover not only a good approximation of coefficients \mathcal{Q} but also the visual weight $\alpha(\nu, r)$ (Eq. 2) that weighted them. A recovered approximation $\hat{\mathcal{Q}}$ with a certain distortion Λ is decoded from the bitstream by the entropy decoding process. The VFWs were not encoded during the entropy encoding process, since it would increase the amount of stored data. A possible solution is to embed these weights $\alpha(\nu, r)$ into $\hat{\mathcal{Q}}$. Thus, our goal is to recover the $\alpha(\nu, r)$ weights only using the information from the bitstream, namely from the Forward quantized coefficients $\hat{\mathcal{Q}}$.

The reduction of the dynamic range is uniformly made by the perceptual quantizer, thus the statistical properties of \mathcal{I} are maintained in $\hat{\mathcal{Q}}$.

Therefore, our hypothesis is that an approximation $\hat{\alpha}(\nu, r)$ of $\alpha(\nu, r)$ can be recovered applying CBPF to \hat{Q} , with the same viewing conditions used in \mathcal{I} . That is, $\hat{\alpha}(\nu, r)$ is the recovered e-CSF. Thus, the perceptual inverse quantizer or the recovered $\hat{\alpha}(\nu, r)$ introduces perceptual criteria to the Inverse Scalar Quantizer [3, Annex J.8] and is given by:

$$\hat{\mathcal{I}} = \begin{cases} \sum_{s=1}^n \sum_{o=v,h,d} \text{sign}(\widehat{\omega}_{s,o}) \frac{\Delta_s \cdot (|\widehat{\omega}_{s,o}^o| + \delta)}{\widehat{\alpha}(\nu, r)} & |\widehat{\omega}_{s,o}| > 0 \\ + (|\widehat{c}_n| + \delta) \cdot \Delta_n, & \\ 0, & \widehat{\omega}_{s,o} = 0 \end{cases} \quad (3)$$

For the sake of showing that the encoded VFWs are approximately equal to the decoded ones, that is $\alpha(\nu, r) \approx \hat{\alpha}(\nu, r)$, we perform the following experiment:

We employ the process shown in Fig. 1(a) for all the images of the CMU [9], CSIQ [10] and IVC [11] Image Databases. We chose for evaluating these assessments the implementation provided in [12], since it is based on the parameters proposed by the author of each indicator. In order to obtain $\hat{\alpha}(\nu, r)$, we measure the lineal correlation between the original $\alpha(\nu, r)$ applied during the F- ρ SQ process and the recovered $\hat{\alpha}(\nu, r)$. Table 2 shows that there is a high similarity between the applied VFW and the recovered one, since their correlation is 0.9849, for gray-scale images, and 0.9840, for color images.

Table 2

CORRELATION BETWEEN $\alpha(\nu, r)$ AND $\hat{\alpha}(\nu, r)$ ACROSS CMU [9], CSIQ [10] AND IVC [11] IMAGE DATABASES.

Image Database	8 bpp gray-scale	24 bpp color
CMU	0.9840	0.9857
CSIQ	0.9857	0.9851
IVC	0.9840	0.9840
Overall	0.9849	0.9844

Fig. 1(b) depicts the PSNR difference (dB) of each color image of the CMU database, that is, the gain in dB of image quality after applying $\hat{\alpha}(\nu, r)$ at $d = 2000$ centimeters to the \hat{Q} images. On average, this gain is about 15 dB. Visual examples of these results are shown by Fig. 2, where the left images are the original images, central images are perceptual quantized images after applying $\alpha(\nu, r)$ and right images are recovered images after applying $\hat{\alpha}(\nu, r)$.

After applying $\hat{\alpha}(\nu, r)$, a visual inspection of these sixteen recovered images show a perceptually lossless quality. We perform the same experiment for gray-scale and color images with $d = 20, 40, 60, 80, 100, 200, 400, 800, 1000$ and 2000 centimeters, in addition to test their objective and subjective image quality by means of the PSNR and MSSIM metrics, respectively, across the CMU(Fig. 3) Image Database.

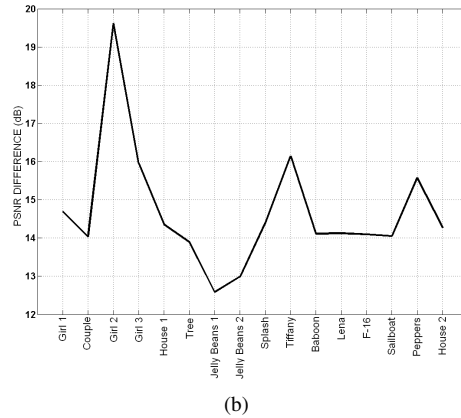
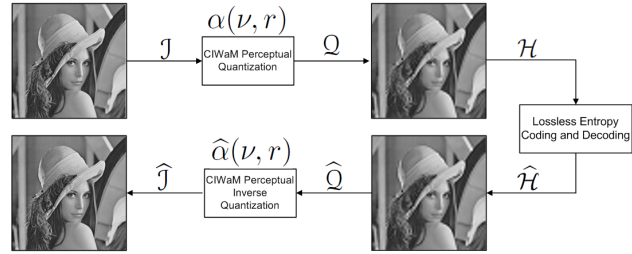
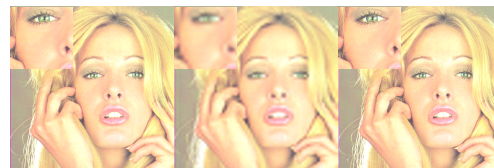


Fig. 1

(A) GRAPHICAL REPRESENTATION OF A WHOLE PROCESS OF COMPRESSION AND DECOMPRESSION. (B)PSNR DIFFERENCE BETWEEN \hat{Q} IMAGE AFTER APPLYING $\alpha(\nu, r)$ AND RECOVERED $\hat{\mathcal{I}}$ AFTER APPLYING $\hat{\alpha}(\nu, r)$ FOR EVERY COLOR IMAGE OF THE CMU DATABASE.



(a) Tiffany



(b) Peppers

Fig. 2

VISUAL EXAMPLES OF PERCEPTUAL QUANTIZATION. LEFT IMAGES ARE THE ORIGINAL IMAGES, CENTRAL IMAGES ARE FORWARD PERCEPTUAL QUANTIZED IMAGES (F- ρ SQ) AFTER APPLYING $\alpha(\nu, r)$ AT $d = 2000$ CENTIMETERS AND RIGHT IMAGES ARE RECOVERED I- ρ SQ IMAGES AFTER APPLYING $\hat{\alpha}(\nu, r)$.

In Figure 3, green functions denoted as F- ρ SQ are the quality metrics of perceptual quantized images after applying

$\alpha(\nu, r)$, while blue functions denoted as $I\text{-}\rho\text{SQ}$ are the quality metrics of recovered images after applying $\hat{\alpha}(\nu, r)$. Thus, either for gray-scale or color images, both PSNR and MSSIM estimations of the quantized image \mathcal{Q} decrease regarding d , the longer d the greater the image quality decline. When the image decoder recovers $\hat{\mathcal{Q}}$ and it is perceptually inverse quantized, the quality barely varies and is close to perceptually lossless, no matter the distance.

5. Experiments and Results

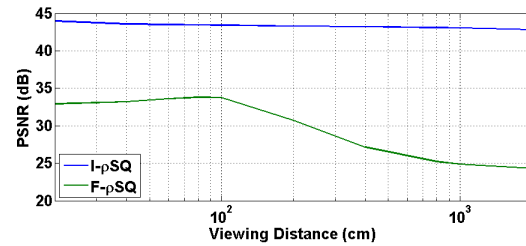
For the sake of comparing the performance between the JPEG2000 [13] and χSET coders, both algorithms are tested according to the process depicted in Fig. 4. First a χSET compression with certain viewing conditions is performed, which gives a compressed image with a particular bit-rate (bpp). Then, a JPEG2000 compression is performed with the same bit-rate. Once both algorithms recover their distorted images, they are compared with some numerical image quality estimators such as: MSSIM [14], PSNR [15], SSIM [16], VIF [17], UQI [18] and WSNR [19].

This experiment is performed across the CMU [9] and IVC [11] Image Databases. Image quality estimations are assessed by the six metrics mentioned before.

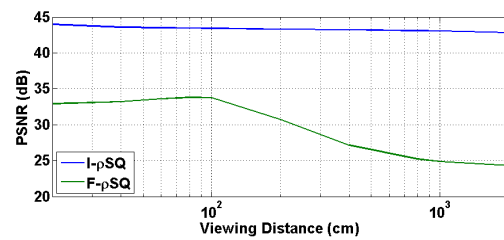
Figs. 5 and 6 show the perceptual quality, estimated by 5(a) MSSIM, 5(c) SSIM, 6(a) UQI, 6(b)VIF and 6(c) WSNR, in addition to the objective quality 5(b) PSNR, of the recovered color images both for JPEG2000(Blue function) and χSET (Green function) as a function of their compression rate. For this experiment, we employ the CMU Image Database and the *Kakadu* implementation for JPEG2000 compression [20]. On the average, a color image compressed at 1.0 bpp (1:24 ratio, stored in 32 KBytes) by JPEG2000 coder has MSSIM=0.9424, SSIM=0.8149, UQI=0.5141, VIF=0.2823 and WSNR=29.2 of perceptual image quality, and PSNR=30.11 of objective image quality, while by χSET has MSSIM=0.9780, SSIM=0.8758, UQI=0.6249, VIF=0.4387, WSNR=35.41 and PSNR=31.84. In Figure 7, we can see these differences when images (a-b)*Lenna*, (c-d)*Girl2* and (e-f)*Tiffany* are compressed at 0.92 bpp, 0.54 bpp and 0.93 bpp, respectively, by JPEG2000 and χSET . For example for these three images, the average difference of MSSIM is 0.0321 in favor of χSET . Therefore, for this image database, χSET has clearly improvement of visual quality than JPEG2000.

6. Conclusions

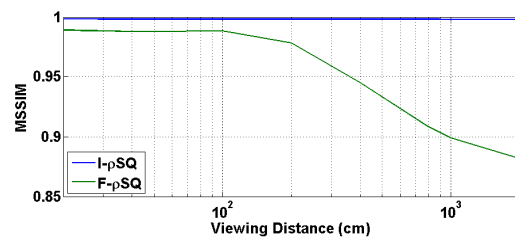
In this work we defined both forward (F- ρSQ) and inverse (I- ρSQ) perceptual quantizer using CBPF. We incorporated it to Hi-SET, proposing the new perceptual image compression system χSET . In order to measure the effectiveness of the perceptual quantization, a performance analysis is done using six image quality assessments such as MSSIM, PSNR, SSIM, UQI, VIF and WSNR, which measured the



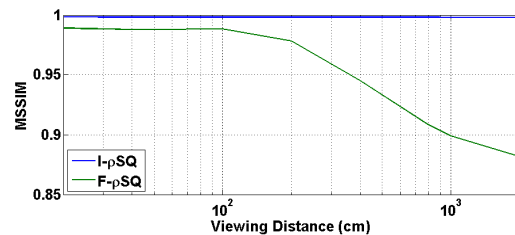
(a) PSNR of Gray-scale



(b) PSNR of Color images



(c) MSSIM of Gray-scale images



(d) MSSIM of Color images

Fig. 3

(A-B) PSNR AND (C-D) MSSIM ASSESSMENTS OF COMPRESSION OF (A & C) GRAY-SCALE (Y CHANNEL) AND (B & D) COLOR IMAGES OF THE CMU IMAGE DATABASE. GREEN FUNCTIONS DENOTED AS F- ρSQ ARE THE QUALITY METRICS OF FORWARD PERCEPTUAL QUANTIZED IMAGES AFTER APPLYING $\alpha(\nu, r)$, WHILE BLUE FUNCTIONS DENOTED AS I- ρSQ ARE THE QUALITY METRICS OF RECOVERED IMAGES AFTER APPLYING $\hat{\alpha}(\nu, r)$.

image quality between reconstructed and original images. The experimental results show that the solely usage of the Forward Perceptual Quantization improves the JPEG2000 compression and image perceptual quality. In addition, when both Forward and Inverse Quantization are applied into Hi-SET, it significantly improves the results regarding the JPEG2000 compression.

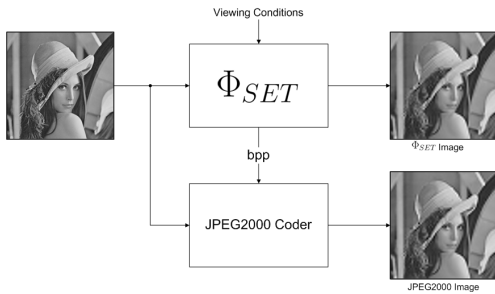
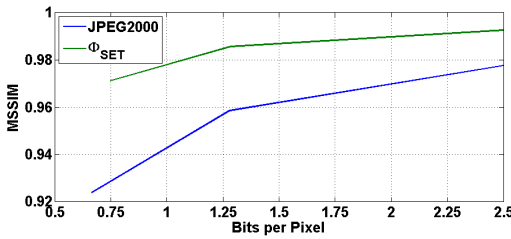
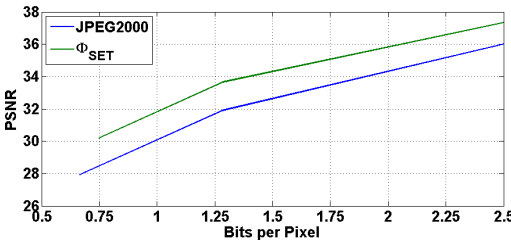


Fig. 4

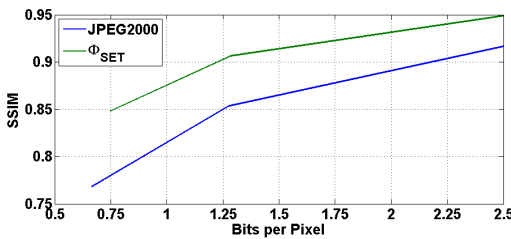
PROCESS FOR COMPARING JPEG2000 AND χ SET. GIVEN SOME VIEWING CONDITIONS A χ SET COMPRESSION IS PERFORMED OBTAINING A PARTICULAR BIT-RATE. THUS, A JPEG2000 COMPRESSION IS PERFORMED WITH SUCH A BIT-RATE.



(a) MSSIM



(b) PSNR

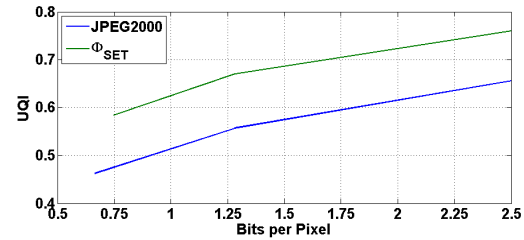


(c) SSIM

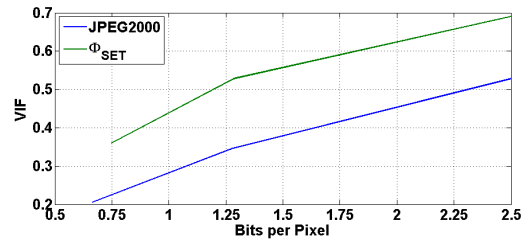
Fig. 5

COMPARISON BETWEEN χ SET (GREEN FUNCTIONS) AND JPEG2000 (BLUE FUNCTIONS) IMAGE CODERS. COMPRESSION RATE VS IMAGE QUALITY ASSESSED BY (A) MSSIM, (B) PSNR AND (C) IN THE CMU IMAGE DATABASE.

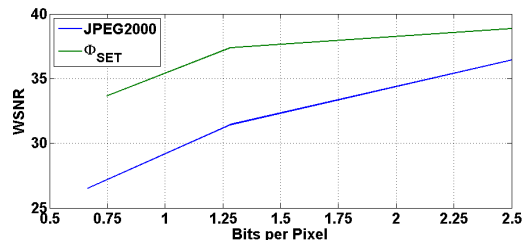
Therefore, we developed a perceptual quantizer algorithm that, in contrast to the JPEG2000 global Frequency weighting, performs a local quantization, that is coefficient-by-



(a) UQI



(b) VIF



(c) WSNR

Fig. 6

COMPARISON BETWEEN χ SET (GREEN FUNCTIONS) AND JPEG2000 (BLUE FUNCTIONS) IMAGE CODERS. COMPRESSION RATE VS IMAGE QUALITY ASSESSED BY (A) UQI, (B) VIF AND (C) WSNR IN THE CMU IMAGE DATABASE.

coefficient. Similarly to JPEG2000, it is not necessary to store the applied weighting for inverse quantizing is because CBPF properties permits to predict perceptual weighting *a posteriori*.

Furthermore, when χ SET is compared with stated-of-the-art perceptual image coders it obtains good results both for objective and subjective images quality.

Acknowledgment

This work is supported by National Polytechnic Institute of Mexico by means of Project No. 20131312 granted by the Academic Secretary and the Committee of Operation and Promotion of Academic Activities (COFAA), National Council of Science and Technology of Mexico by means of Project No. 204151/2013, LABEX Σ -LIM France, Coimbra Group Scholarship Programme granted by University of Poitiers and Region of Poitou-Charentes, France.



Fig. 7

EXAMPLE OF RECOVERED COLOR IMAGES *Lenna*, *Girl2* AND *Tiffany* OF THE CMU IMAGE DATABASE COMPRESSED AT (A AND B) 0.92 BPP, ((B AND C) 0.54 BPP AND (C AND D) 0.93 BPP, RESPECTIVELY. JPEG2000 IMAGES ARE COMPRESSED USING TABLE 1 AND $s = 3$.

References

- [1] W. M. Goodall, "Television by pulse code modulation," *Bell Syst. Techn. J.*, vol. 28, p. 33-49, Jan. 1951.
- [2] M. Nadenau, J. Reichel, and M. Kunt, "Visually improved image compression by combining a conventional wavelet-codec with texture modeling," *IEEE Trans. Image Process.*, vol. 11, no. 11, pp. 1284-1294, Nov. 2002.
- [3] M. Boliek, C. Christopoulos, and E. Majani, *Information Technology: JPEG2000 Image Coding System*, JPEG 2000 Part I final committee draft version 1.0 ed., ISO/IEC JTC1/SC29 WG1, JPEG 2000, April 2000.
- [4] X. Otazu, C. Párraga, and M. Vanrell, "Toward a unified chromatic induction model," *Journal of Vision*, vol. 10(12), no. 6, 2010.
- [5] K. Mullen, "The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings," *Journal of Physiology*, vol. 359, pp. 381-400, 1985.
- [6] J. Moreno and X. Otazu, "Image coder based on Hilbert Scanning of Embedded quadTrees," *IEEE Data Compression Conference*, p. 470, March 2011.
- [7] —, "Image coder based on Hilbert Scanning of Embedded quadTrees," June 2011, under review in *Digital Signal Processing*.
- [8] —, "Image coder based on Hilbert Scanning of Embedded quadTrees: An introduction of Hi-SET coder," *IEEE International Conference on Multimedia and Expo*, July 2011.
- [9] S. I. P. I. of the University of Southern California. (1997) The USC-SIPI image database. Signal and Image Processing Institute of the University of Southern California. [Online]. Available: <http://sipi.usc.edu/database/>
- [10] E. C. Larson and D. M. Chandler, "Most apparent distortion: a dual strategy for full-reference image quality assessment," in *Proc. SPIE*, vol. 742, 2009.
- [11] P. le Callet and F. Atrousseau, "Subjective quality assessment IRC-CyN/IVC database," 2005, <http://www.irccyn.ec-nantes.fr/ivcdb/>.
- [12] C. U. V. C. Laboratory. (2010) MeTriX MuX visual quality assessment package, available at http://foulard.ece.cornell.edu/gaubatz/metrix_mux/. Cornell University Visual Communications Laboratory.
- [13] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, ser. ISBN: 0-7923-7519-X. Kluwer Academic Publishers, 2002.
- [14] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers.*, vol. 2, 2003, pp. 1398 - 1402 Vol.2.
- [15] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800-801, 2008.
- [16] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430 -444, feb. 2006.
- [17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600 -612, april 2004.
- [18] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81-84, 2002.
- [19] T. Mitsa and K. Varkur, "Evaluation of contrast sensitivity functions for formulation of quality measures incorporated in halftoning algorithms," *IEEE International Conference on Acustics, Speech and Signal Processing*, vol. 5, pp. 301-304, 1993.
- [20] D. Taubman. (2010, July) Kakadu software. [Online]. Available: <http://www.kakadusoftware.com/>

Night Color Image Enhancement via Statistical Law and Retinex

Huaxia Zhao*, Chuangbai Xiao, Hongyu Zhao

College of Computer Science and Technology, Beijing University of Technology, Beijing, China

Abstract - Due to the uneven distribution of light at night, the quality of the night color image is usually unsatisfactory. To solve this problem, in this paper, we propose a new statistical method based on the retinex. The algorithm analyzes the transformation relationship between the nighttime image and illumination image by the algorithm of Michael Elad and MSRCR algorithm. Through this transformation, we can accurately and quickly get the illumination image. Then, we can get the resulting image successfully based on the retinex. Our algorithm can greatly enhance the image contrast and brightness, recover image details, eliminate the “halo effect” efficiently, and accelerate the computational speed remarkably. Experiments on different nighttime images demonstrate the effectiveness of our approach.

Keywords: statistical law; image enhancement; night color image

1 Introduction

The night color image enhancement is of great importance in both the computational photography and computer vision. First, it can effectively increase the visibility and surrealism of the scene. Second, artificial illumination light distributes unevenly at night leading to weakening the quality of monitoring photos and increasing the difficulty of surveillance. Thus, the night color image enhancement can promote the video surveillance. Last, the input images of most computer vision algorithms (e.g., the photometric analysis algorithm) are daytime images. Thus, the night color image enhancement can increase the scope of such algorithms by enhancing nighttime images.

However, the night color image enhancement is a challenging task. Currently, the main techniques for the night image enhancement are the image fusion and image enhancement. Image fusion techniques include two categories: one is the fusion of the nighttime image and visible image [1, 2] and another is the fusion of the nighttime image and infrared image [3, 4]. These methods require multiple different spectral images collected in the same scene and have high computational complexity. The main techniques for the image enhancement include contrast stretching, slicing, histogram equalization, and some algorithms based on the retinex [5-11], etc. Of all those algorithms, the algorithm based on the retinex has acceptable results, but it will produce the “halo effect” and high time

complexity.

In this paper, we propose a novel algorithm for enhancing the night color image based on the statistical law and the retinex. We assume that there is a transformation on the brightness components of the pixel values between the nighttime image and illumination image. Therefore, through this transformation, we can accurately and quickly get the illumination image. Then, we can get the resulting image successfully based on the retinex. The resulting image retains image details and exhibits higher brightness, so that the overall image looks more harmonious and natural. Our algorithm is simpler and faster compared to the other algorithms.

The rest of this paper is organized as follows. In section II, overview of retinex theory is given. In Section III, the proposed algorithm is described in detail, which contains two parts: analysis of the transformation law and enhancing the nighttime image. Finally, the experimental results are presented to demonstrate the efficiency of the proposed algorithm in section IV. The conclusion is in Section V.

2 Overview of retinex theory

The Retinex theory deals with the removal of unfavorable illumination effects from a given image. A commonly assumed model suggests that any given image S is the pixel-wise multiplication of two images, the reflection image R and the illumination image L . This model is given in Eq. (1):

$$S = R \cdot L \quad (1)$$

Therefore, if we can get the illumination image, we can quickly get the reflection image. In the actual calculation, a look-up-table log operation transfers this multiplication into an addition, resulting with $s = \log(S) = \log(L) + \log(R) = l + r$.

3 Analyzing the transformation law and enhancing the nighttime image

We first transform the original RGB space to HSV space, because processing the color image directly in RGB space will lead to color distortion. The HSV space is closer to human visual perception in color perception. Our transformation law is only used in brightness component of HSV space.

Currently, most algorithms often use the filtering method to estimate the illumination image, and achieve good results. In this paper, we use the processing results of some algorithms (the algorithm of Michael Elad [12] and MSRCR [13,14]) as illumination images. Through these two algorithms, we get three images. One is the nighttime image, and the other two are the corresponding illumination images. Their brightness components of HSV space are denoted as L, M and N. For analyzing the transformation law, we get pixels which value is i (0-255) from L. Then, we have a set of coordinates through known pixels. In the same coordinates, we get two sets of pixel values from M and N. The average (j, k) of these two sets are the corresponding value to i . Fig.1 displays the correspondence between i and j, k . In order to facilitate observation, we add a linear which is $y=x$.

By observing Fig.1, we find the curve of MSRCR on the figure can be represented by a circular arc. But it is too close to the linear which is $y = x$ resulting in that the enhanced image is too bright and loses details seriously. The curve of Michael Elad is roughly like a circular arc except a small part. It is the reason that the resulting images processed by the algorithm of Michael Elad have a stronger noise. Overall, we can use a circular arc to represent the relationship between the input image and illumination image. Obviously, the fitting circular arc should pass the point $(255,255)$. In order to facilitate the calculation, we use two parameters to represent the circular arc. One parameter is x -coordinate (x_0) of the circular center. Another is the intersection $(0, \lambda)$ of arc and y positive axle. According to the nature of the circle, the y -coordinate (y_0) of the circular center can be expressed as the following:

$$y_0 = \frac{255^2 - \lambda^2 / 2 - 255 * x_0}{255 - \lambda} \quad (2)$$

The radius(r) of the circular arc can be expressed as the following:

$$r = \sqrt{(x_0 - 255)^2 + (y_0 - 255)^2} \quad (3)$$

The circular arc can be expressed as the following:

$$y = \sqrt{r^2 - (x - x_0)^2} + y_0 \quad (4)$$

Using Eq. (4), we can get the illumination image directly and quickly. The circular arc is shown in the Fig.2.

According to the Eq. (4), we can know that the circular arc takes two parameters called x_0 and λ . The smaller x_0 is, the more obvious the brightness enhancement is. The greater λ is, the more obvious the brightness enhancement is. It can be seen that the darker the input image is, the smaller x_0 is and the greater λ is, and vice versa. Therefore, we assign the average pixel value of the source image to λ . x_0 is represented by the following formula:

$$x_0 = \max(127, \text{round}(6000 * \exp^{-\lambda/30})) \quad (5)$$

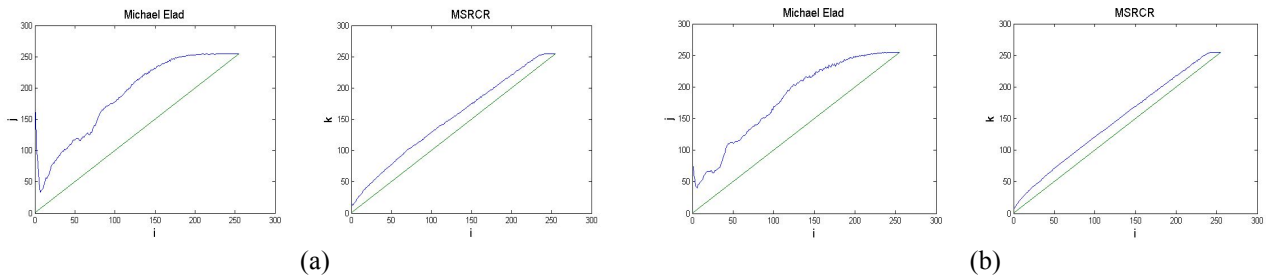
Where $\text{round}()$ is the rounding function which is used to improve the algorithm speed. The minimum size of x_0 is limited to 127 to prevent image distortion. Therefore, we have got a no-argument function to show the relationship between the original image and the corresponding illumination image.

The algorithm is given below:

- 1) Transform the original RGB data to the HSV data.
- 2) Get the illumination image using Eq. (4, 5).
- 3) Enhance the nighttime image through the retinex theory and the obtained illumination image.
- 4) Transform the HSV data to the RGB data and show the enhanced image.

4 Comparison and results

In this section, the proposed algorithm of this paper is compared with the algorithm of Michael Elad and MSRCR algorithm. Fig.3-7 show the experimental results of five scenes.(a) is the original image.(b) shows the enhanced image by the algorithm of Michael Elad.(c) is obtained by MSRCR algorithm.(d) displays the enhanced image by the proposed algorithm in this paper.



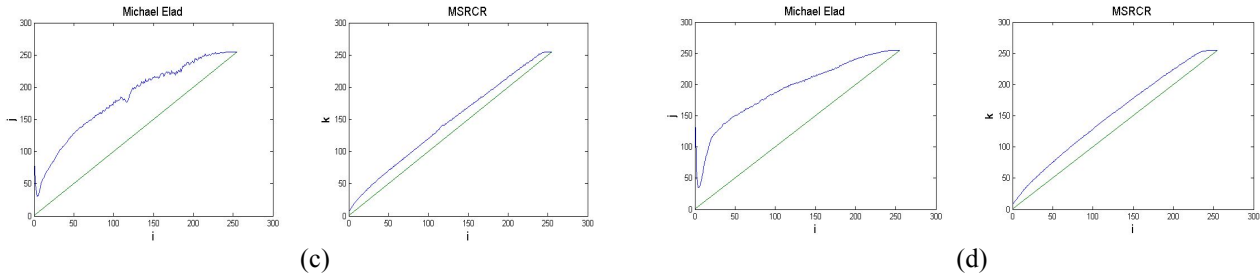


Fig.1. the corresponding graphs of pixel values of source images and illumination images obtained by the algorithm of Michael Elad and MSRCR algorithm. (a), (b), (c) and (d) are the processing results of four different pictures.

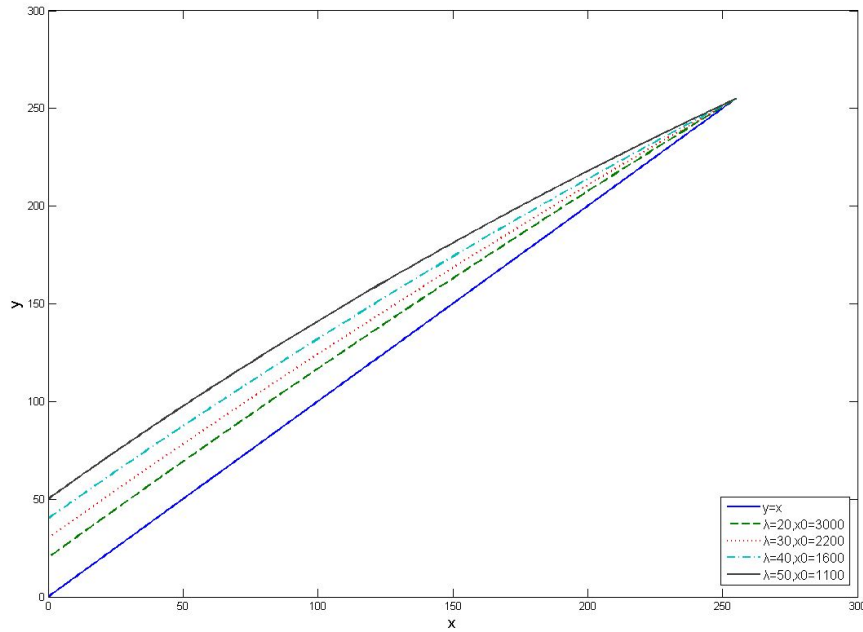


Fig.2. Fitting circular arc

It can be clearly seen from the figures that the three algorithms all have a certain enhancement effect on the nighttime images. The processing results of MSRCR algorithm are too bright leading to atomization phenomenon and loss of details, as shown in Fig.4 (c), Fig.7 (c) and Fig.8 (c). The processing results of Michael Elad's algorithm tend to produce excessive sharpening phenomenon in highlighting edges and the "halo artifacts" phenomenon which is shown in Fig.6 (b) and Fig.7 (b). Moreover, it often leads to noisy amplification in dark areas, as shown in Fig.4 (b), which is mainly due to the unsmooth curve (shown in Fig.1) between source images and illumination images. Compared with the two algorithms, the proposed algorithm can better recover details, eliminates the "halo effect" and suppress noise. Moreover,

the results of our algorithm look more harmonious and natural.

For a more definite description of the experimental results, this paper also uses objective evaluation criteria to test the effectiveness of our algorithm. We examine our algorithm in mean value, standard deviation and time-consuming (Computer configuration: CPU: Pentium(R) 3.00GHz; Memory: 3.00GB; Programming Language: Matlab). The image mean reflects the brightness level of the image; the standard deviation reflects the image contrast; the time-consuming reflects the time complexity of the algorithm. The results are shown in Table 1-5.



Fig.3. Scene 1

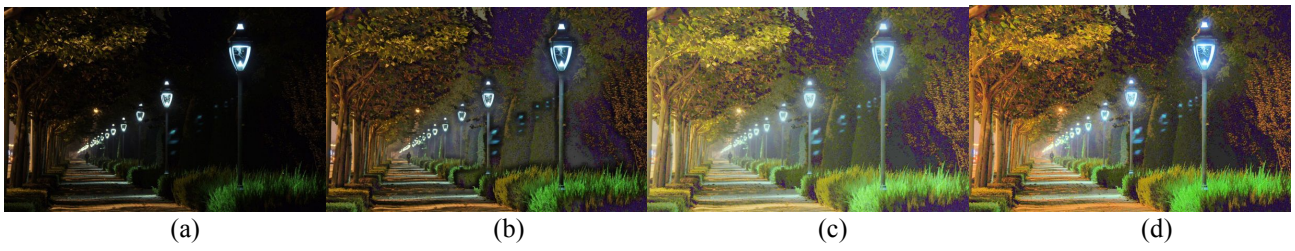


Fig.4. Scene 2

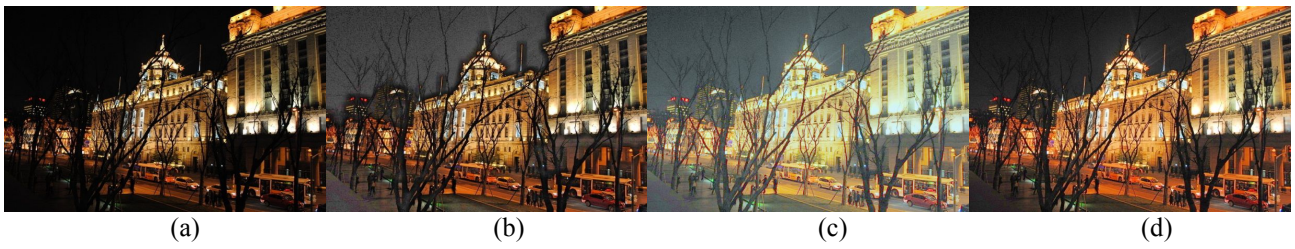


Fig.5. Scene 3



Fig.6. Scene 4



Fig.7. Scene 5

Table 1.Criteria of Assessment of Fig.3

	mean	standard deviation	time-consuming(s)
Source image	32.290669	37.976650	
Michael Elad	74.522833	34.043863	125.203029
MSRCR	122.023486	33.550094	8.073667
proposed algorithm	92.0300	38.6720	1.246996

Table 2.Criteria of Assessment of Fig.4

	mean	standard deviation	time-consuming(s)
Source image	19.074728	28.879643	
Michael Elad	54.979484	32.582415	174.404951
MSRCR	105.600284	41.911008	14.030690
proposed algorithm	87.0683	46.6404	1.724475

Table 3.Criteria of Assessment of Fig.5

	mean	standard deviation	time-consuming(s)
Source image	42.871070	57.089267	
Michael Elad	74.523826	51.653054	177.822975
MSRCR	125.407817	52.732457	11.619577
proposed algorithm	81.8586	61.3861	1.39663

Table 4.Criteria of Assessment of Fig.6

	mean	standard deviation	time-consuming(s)
Source image	43.9788	37.4953	
Michael Elad	85.4186	28.9512	177.513603
MSRCR	145.5552	27.6842	13.096727
proposed algorithm	94.7658	32.4271	1.450654

Table 5.Criteria of Assessment of Fig.7

	mean	standard deviation	time-consuming(s)
Source image	37.186166	32.700994	
Michael Elad	87.693053	47.901984	306.097388
MSRCR	141.732310	45.151938	27.631109
proposed algorithm	101.8945	55.9020	2.418071

As can be seen in Table 1-5, in terms of the mean, MSRCR algorithm has the most significant effect of improving mean, but the enhanced images are too bright overall to protect details. Compared with the

algorithm of Michael Elad, the proposed algorithm has a better enhancing effect of the mean which displays the brightness of the whole picture is consistent with human visual perception. In terms of standard deviation, the proposed algorithm is better than the other two algorithms. The proposed algorithm has remarkable enhancement of the image contrast and significant effect of image detail recovery. In terms of time-consuming, the proposed algorithm has lower time complexity than the other two algorithms. Moreover, our algorithm does not need manual control parameters increasing the adaptability of our algorithm.

5 The conclusion

In this paper we present an effective algorithm for enhancing the nighttime image. In our algorithm, we propose a statistical law to present the relation of the original image and illumination image. Using this statistical law and retinex theory, we can accurately and quickly get the resulting image. The algorithm is validated through subjective and objective evaluation, which shows it can eliminate the “halo effect”, enhance the image contrast, recover image details and have low time complexity. In summary, our algorithm is effective to complete the challenging task of enhancing the nighttime image.

6 References

- [1] RASKAR R, ILIE A, YU Jingyi. Image fusion for context enhancement and video surrealism: Proceedings NPAR 2004 - 3rd International Symposium on Non-Photorealistic Animation and Rendering, 2004[C]. Annecy, France: Association for Computing Machinery, 2004: 85-93.
- [2] YAMASAKI A, TAKAUJI H, KANEKO S, et al. Denighting: Enhancement of nighttime images for a surveillance camera: 2008 19th International Conference on Pattern Recognition, ICPR 2008, 2008[C]. Tampa, FL, United states: Institute of Electrical and Electronics Engineers Inc, 2008.
- [3] ZHANG Xiaopeng, SIM T, MIAO Xiaoping. Enhancing photographs with near infrared images: 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008[C]. Anchorage, AK, United states: Inst. of Elec. and Elec. Eng. Computer Society, 2008.
- [4] ZHUO Shaojie, ZHANG Xiaopeng, MIAO Xiaoping, et al. Enhancing low light images using near infrared flash images: 2010 17th IEEE International Conference on Image Processing, ICIP 2010,2010[C]. Hong Kong, Hong kong: IEEE Computer Society, 2010: 2537-2540.
- [5] BRAINARD D, WANDELL B. Analysis of the retinex theory of color vision [J]. Journal of the Optical Society of America, 1986. 3: 1651-1661.

- [6] MCCANN, JOHN. Lessons learned from Mondrians applied to real images and color gamuts: Final Program and Proceedings of the 7th IS and T/SID Color Imaging Conference: Color Science, Systems and Applications, 1999[C]. Scottsdale, AZ, United states: Society for Imaging Science and Technology, 1999:1-8.
- [7] FUNT B, CIUREA F, MCCANN J. Retinex in Matlab: Final Program and Proceedings of the 8th IS and T/SID Color Imaging Conference: Color Science, Systems and Applications, 2000[C]. Scottsdale, AZ, United states: Society for Imaging Science and Technology, 2000:112-121.
- [8] JOBSON D J, RAHMAN Z, WOODDELL G A. Properties and performance of a center/surround retinex [J]. IEEE Transactions on Image Processing, 1997. 6(3): 451-462.
- [9] RAHMAN Z, JOBSON D J, WOODDELL G A. Multi-scale retinex for color image enhancement: Proceedings of the 1996 IEEE International Conference on Image Processing, ICIP'96. 1996[C]. Lausanne, Switz: IEEE, 1996: 1003-1006.
- [10] TOMASI C, MANDUCHI R. Bilateral filtering for gray and color images: Proceedings of the 1998 IEEE 6th International Conference on Computer Vision, 1998[C]. Bombay, India: IEEE, 1998: 839-846.
- [11] MEYLAN L, SUSSTRUNK S. High dynamic range image rendering with a retinex-based adaptive filter [J]. IEEE Transactions on Image Processing, 2006. 15(9): 2820-2830.
- [12] Michael Elad. Retinex by two bilateral filters. In Proceedings of the scale-space conference, 2005 9(7):217-229.
- [13] JOBSON D J, RAHMAN Z, WOODDELL G A. Multiscale retinex for bridging the gap between color images and the human observation of scenes[J]. IEEE Transactions on Image Processing, 1997. 6(7): 965-976.
- [14] RAHMAN Z, JOBSON D J, WOODDELL G A. Retinex processing for automatic image enhancement [J]. Journal of Electronic Imaging, 2004. 13(1): 100-110.

ρ SQ: Image Quantizer based on Contrast Band-Pass Filtering

Jaime Moreno^{†‡}, Oswaldo Morales[†], and Ricardo Tejada[†]

[†]Superior School of Mechanical and Electrical Engineers, National Polytechnic Institute of Mexico, IPN Avenue, Lindavista, Mexico City, 07738, Mexico.

[‡]Signal, Image and Communications Department, University of Poitiers, Poitiers, 30179, France.
e-mail:jmorenoe@ipn.mx

Abstract—*The aim of this work is to explain how to apply perceptual criteria in order to define a perceptual forward and inverse quantizer. We present its application to the Hi-SET coder. Our approach consists in quantizing wavelet transform coefficients using some of the human visual system behavior properties. Taking in to account that noise is fatal to image compression performance, because it can be both annoying for the observer and consumes excessive bandwidth when the imagery is transmitted. Perceptual quantization reduces unperceivable details and thus improve both visual impression and transmission properties. The comparison between JPEG2000 coder and the combination of Hi-SET with the proposed perceptual quantizer (χ SET) shows that the latter is not favorable in PSNR than the former, but the recovered image is more compressed (less bit-rate) at the same or even better visual quality measured with well-know image quality metrics, such as MSSIM, UQI or VIF, for instance.*

Keywords: *Human Visual System, Contrast Sensitivity Function, Perceived Images, Wavelet Transform, Peak Signal-to-Noise Ratio, No-Reference Image Quality Assessment, JPEG2000.*

1. Introduction

Digital image compression has been a research topic for many years and a number of image compression standards has been created for different applications. The JPEG2000 is intended to provide rate-distortion and subjective image quality performance superior to existing standards, as well as to supply functionality [1]. However, JPEG2000 does not provide the most relevant characteristics of the human visual system, since for removing information in order to compress the image mainly information theory criteria are applied. This information removal introduces artifacts to the image that are visible at high compression rates, because of many pixels with high perceptual significance have been discarded.

Hence, it is necessary an advanced model that removes information according to perceptual criteria, preserving the pixels with high perceptual relevance regardless of the numerical information. The Chromatic Induction Wavelet Model presents some perceptual concepts that can be suitable for

it. Both CBPF and JPEG2000 use wavelet transform. CBPF uses it in order to generate an approximation to how every pixel is perceived from a certain distance taking into account the value of its neighboring pixels. By contrast, JPEG2000 applies a perceptual criteria for all coefficients in a certain spatial frequency independently of the values of its surrounding ones. In other words, JPEG2000 performs a global transformation of wavelet coefficients, while CBPF performs a local one.

CBPF attenuates the details that the human visual system is not able to perceive, enhances those that are perceptually relevant and produces an approximation of the image that the brain visual cortex perceives. At long distances the lack of information does not produce the well-known compression artifacts, rather it is presented as a softened version, where the details with high perceptual value remain (for example, some edges).

2. JPEG2000 Global Visual Frequency Weighting

In JPEG2000, only one set of weights is chosen and applied to wavelet coefficients according to a particular viewing condition (100, 200 or 400 dpi's) with fixed visual weighting[1, Annex J.8]. This viewing condition may be truncated depending on the stages of embedding, in other words at low bit rates, the quality of the compressed image is poor and the detailed features of the image are not available since at a relatively large distance the low frequencies are perceptually more important.

The table 1 specifies a set of weights which was designed for the luminance component based on the CSF value at the mid-frequency of each spatial frequency. The viewing distance is supposed to be 4000 pixels, corresponding to 10 inches for 400 dpi print or display. The weight for *LL* is not included in the table, because it is always 1. Levels 1, 2, . . . , 5 denote the spatial frequency levels in low to high frequency order with three spatial orientations, *horizontal*, *vertical* and *diagonal*.

Table 1

RECOMMENDED JPEG2000 FREQUENCY (s) WEIGHTING FOR 400 DPI'S
($s = 1$ IS THE LOWEST FREQUENCY WAVELET PLANE).

s	horizontal	vertical	diagonal
1	1	1	1
2	1	1	0.731 668
3	0.564 344	0.564 344	0.285 968
4	0.179 609	0.179 609	0.043 903
5	0.014 774	0.014 774	0.000 573

3. Perceptual Forward Quantization

3.1 Methodology

Quantization is the only cause that introduces distortion into a compression process. Since each transform sample at the perceptual image \mathcal{I}_ρ (from Eq. ??) is mapped independently to a corresponding step size either Δ_s or Δ_n , thus \mathcal{I}_ρ is associated with a specific interval on the real line. Then, the perceptually quantized coefficients \mathcal{Q} , from a known viewing distance d , are calculated as follows:

$$\mathcal{Q} = \sum_{s=1}^n \sum_{o=v,h,d} \text{sign}(\omega_{s,o}) \left[\frac{|\alpha(\nu, r) \cdot \omega_{s,o}|}{\Delta_s} \right] + \left\lfloor \frac{c_n}{\Delta_n} \right\rfloor \quad (1)$$

Unlike the classical techniques of Visual Frequency Weighting (VFW) on JPEG2000, which apply one CSF weight per sub-band [1, Annex J.8], Perceptual Quantization using CBPF (ρ SQ) applies one CSF weight per coefficient over all wavelet planes $\omega_{s,o}$. In this section we only explain Forward Perceptual Quantization using CBPF (F- ρ SQ). Thus, Equation 1 introduces the perceptual criteria of Perceptual Images to each quantized coefficient of Equation of Dead-zone Scalar Quantizer. A normalized quantization step size $\Delta = 1/128$ is used, namely the range between the minimal and maximal values at \mathcal{I}_ρ is divided into 128 intervals. Finally, the perceptually quantized coefficients are entropy coded, before forming the output code stream or bitstream.

3.2 Experimental Results applied to JPEG2000

The Perceptual quantizer F- ρ SQ in JPEG2000 is tested on all the color images of the *Miscellaneous volume* of the University of Southern California Image Data Base[2]. The data sets are eight 256×256 pixel images (Fig. ??) and eight 512×512 pixel images (Fig. ??), but only visual results of the well-known images *Lena*, *F-16* and *Baboon* are depicted, which are 24-bit color images and 512×512 of resolution. The CBPF model is performed for a 19 inch monitor with 1280 pixels of horizontal resolution at 50 centimeters of viewing distance. The software used to obtain a JPEG2000 compression for the experiment is *JJ2000*[3].

Figure 1 shows the assessment results of the average performance of color image compression for each bit-plane using a Dead-zone Uniform Scalar Quantizer (SQ, function

with heavy dots), and it also depicts the results obtained when applying F- ρ SQ(function with heavy stars).

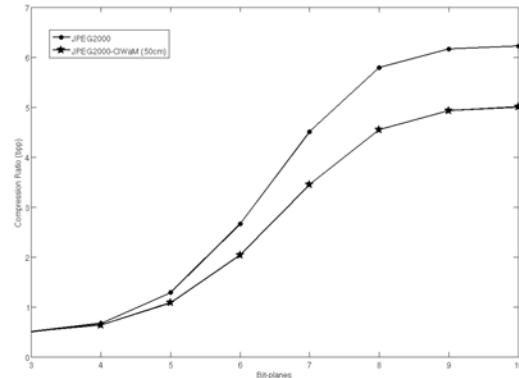


Fig. 1

JPEG2000 COMPRESSION RATIO (BPP) AS A FUNCTION OF BIT-PLANE. FUNCTION WITH HEAVY DOTS SHOWS JPEG2000 ONLY QUANTIZED BY THE DEAD-ZONE UNIFORM SCALAR MANNER. WHILE FUNCTION WITH HEAVY STARS SHOWS JPEG2000 PERCEPTUALLY PRE-QUANTIZED BY F- ρ SQ.

Using CBPF as a method of forward quantization, achieves better compression ratios than SQ with the same threshold, obtaining better results at the highest bit-planes, since CBPF reduces unperceivable features. Figure 2 shows the contribution of F- ρ SQ in the JPEG2000 compression ratio, for example, at the eighth bit-plane, CBPF reduces 1.2423 bits per pixel than the bit rate obtained by SQ, namely in a 512×512 pixel color image, CBPF estimates that 39.75KB of information is perceptually irrelevant at 50 centimeters.

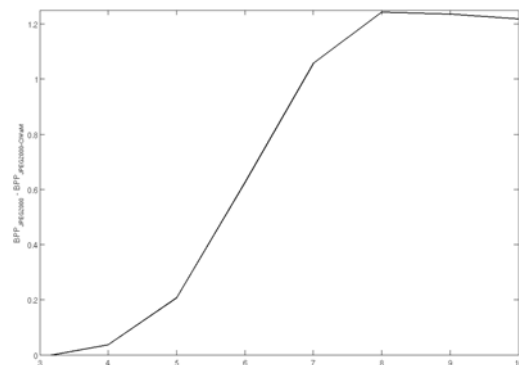
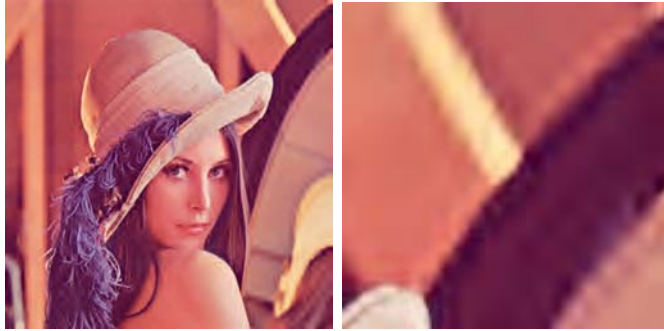


Fig. 2

THE BIT-RATE DECREASE BY EACH BIT-PLANE AFTER APPLYING F- ρ SQ ON THE JPEG2000 COMPRESSION.

Both Figure 3 and 4 depict examples of recovered images compressed at 0.9 and 0.4 bits per pixel, respectively, by means of JPEG2000 (a) without and (b) with F- ρ SQ. Also these figures show that the perceptual quality of images forward quantized by ρ SQ is better than the objective one.



(a) JPEG2000 PSNR=31.19 dB.

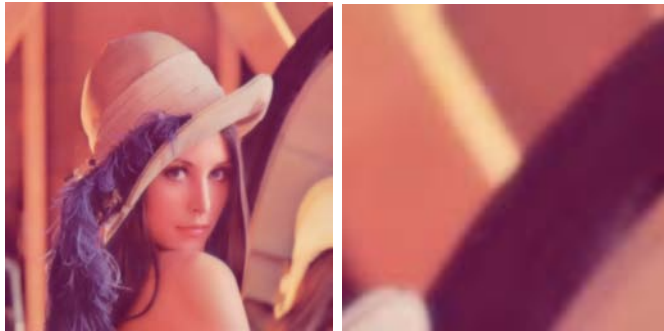
(b) JPEG2000-F- ρ SQ PSNR=27.57 dB.

Fig. 3

EXAMPLES OF RECOVERED IMAGES OF LENA COMPRESSED AT 0.9 BPP.



(a) JPEG2000 PSNR=25.12 dB.

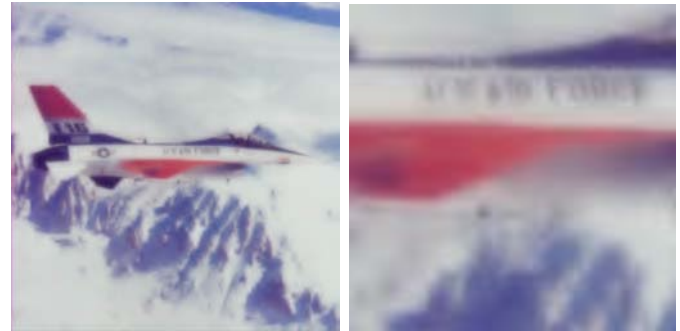
(b) JPEG2000-F- ρ SQ PSNR=24.57 dB.

Fig. 4

EXAMPLES OF RECOVERED IMAGES OF F-16 COMPRESSED AT 0.4 BPP.

Figure 5 shows examples of recovered images of *Baboon* compressed at 0.59, 0.54 and 0.45 bits per pixel by means of JPEG2000 (a) without and (b and c) with F- ρ SQ. In Fig. 5(a) PSNR=26.18 dB and in Fig. 5(b) PSNR=26.15 dB but a perceptual metrics like WSNR [4], for example, assesses that it is equal to 34.08 dB. Therefore, the recovered image Forward quantized by ρ SQ is perceptually better than the one only quantized by a SQ. Since the latter produces more compression artifacts, the ρ SQ result at 0.45 bpp (Fig. 5(c)) contains less artifacts than SQ at 0.59 bpp. For example the *Baboon's* eye is softer and better defined using F- ρ SQ and it additionally saves 4.48 KB of information.

4. Perceptual Inverse Quantization

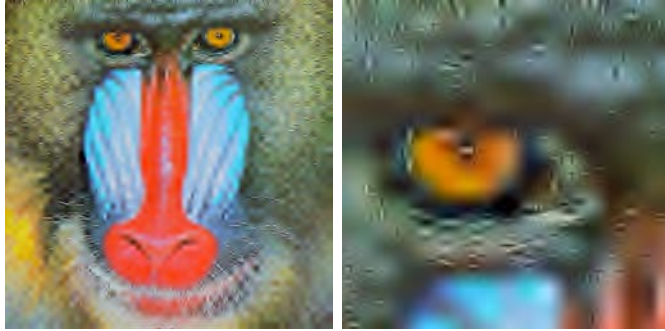
The proposed Perceptual Quantization is a generalized method, which can be applied to wavelet-transform-based image compression algorithms such as EZW, SPIHT, SPECK or JPEG2000. In this work, we introduce both forward (F- ρ SQ) and inverse perceptual quantization (I- ρ SQ) into the *Hi-SET* coder. This process is shown in the green blocks of Fig. 6. An advantage of introducing ρ SQ is to maintain the embedded features not only of *Hi-SET* algorithm but also of any wavelet-based image coder. Thus, we call CBPF Perceptual Quantization + *Hi-SET* = *cHi-SET* or χ SET.

Both JPEG2000 and χ SET choose their VFWs according to a final viewing condition. When JPEG2000 modifies the

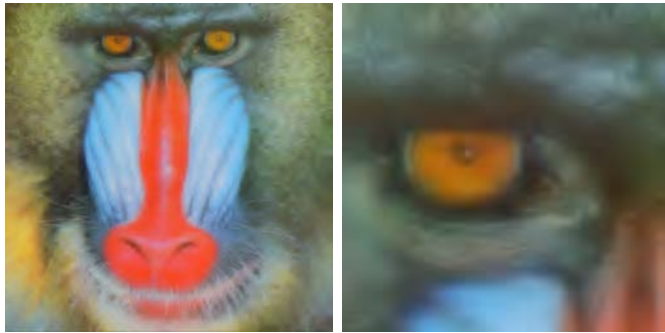
quantization step size with a certain visual weight, it needs to explicitly specify the quantizer, which is not very suitable for embedded coding. While χ SET neither needs to store the visual weights nor to necessarily specify a quantizer in order to keep its embedded coding properties.

The main challenge underlies in to recover not only a good approximation of coefficients \hat{Q} but also the visual weight $\alpha(\nu, r)$ (Eq. 1) that weighted them. A recovered approximation \hat{Q} with a certain distortion Λ is decoded from the bitstream by the entropy decoding process. The VFWs were not encoded during the entropy encoding process, since it would increase the amount of stored data. A possible solution is to embed these weights $\alpha(\nu, r)$ into \hat{Q} . Thus, our goal is to recover the $\alpha(\nu, r)$ weights only using the information from the bitstream, namely from the Forward quantized coefficients \hat{Q} .

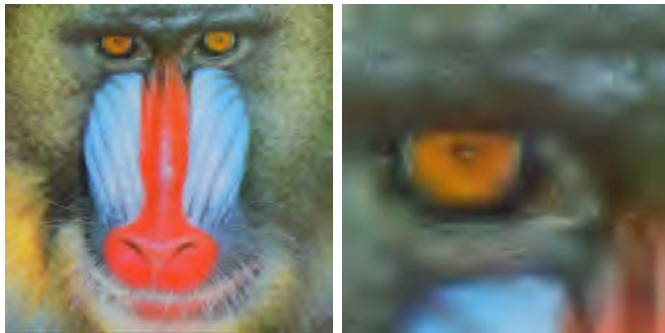
Therefore, our hypothesis is that an approximation $\hat{\alpha}(\nu, r)$ of $\alpha(\nu, r)$ can be recovered applying CBPF to \hat{Q} , with the same viewing conditions used in \mathcal{I} . That is, $\hat{\alpha}(\nu, r)$ is the recovered e-CSF. Thus, the perceptual inverse quantizer or the recovered $\hat{\alpha}(\nu, r)$ introduces perceptual criteria to Inverse Scalar Quantizer and is given by:



(a) JPEG2000 compressed at 0.59 bpp.



(b) JPEG2000-F-ρSQ compressed at 0.54 bpp.



(c) JPEG2000-F-ρSQ compressed at 0.45 bpp.

Fig. 5

EXAMPLES OF RECOVERED IMAGES OF BABOON.



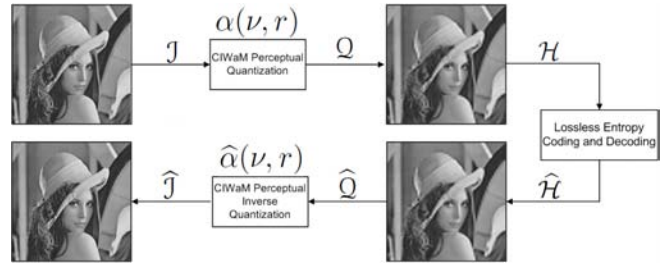
Fig. 6

THE χ SET IMAGE COMPRESSION ALGORITHM. GREEN BLOCKS ARE THE F-ρSQ AND I-ρSQ PROCEDURES.

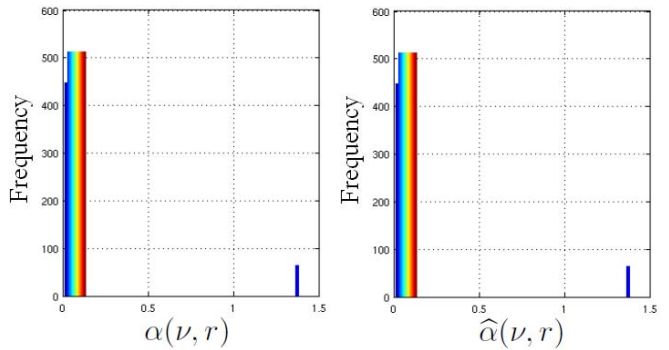
$$\hat{\mathcal{I}} = \begin{cases} \sum_{s=1}^n \sum_{o=v,h,d} \text{sign}(\widehat{\omega}_{s,o}) \frac{\Delta_s \cdot (|\widehat{\omega}_{s,o}| + \delta)}{\hat{\alpha}(\nu, r)} + (\widehat{c}_n + \delta) \cdot \Delta_n & |\widehat{\omega}_{s,o}| > 0 \\ 0, & \widehat{\omega}_{s,o} = 0 \end{cases} \quad (2)$$

For the sake of showing that the encoded VFWS are approximately equal to the decoded ones, that is $\alpha(\nu, r) \approx \hat{\alpha}(\nu, r)$, we perform two experiments.

Experiment 1: Histogram of $\alpha(\nu, r)$ and $\hat{\alpha}(\nu, r)$. The process of this short experiment is shown by Figure 7. Figure 7(a) depicts the process for obtaining losslessly both Encoded and Decoded visual weights for the 512×512 *Lena* image, channel *Y* at 10 meters. While Figures 7(b) and 7(c) shows the frequency histograms of $\alpha(\nu, r)$ and $\hat{\alpha}(\nu, r)$, respectively. In both graphs, the horizontal axis represents the sort of VFW variations, whereas the vertical axis represents the number of repetitions in that particular VFW. The distribution in both histograms is similar and they have the same shape.



(a)



(b)

(c)

Fig. 7

(A) GRAPHICAL REPRESENTATION OF A WHOLE PROCESS OF COMPRESSION AND DECOMPRESSION. HISTOGRAMS OF (B) $\alpha(\nu, r)$ AND (C) $\hat{\alpha}(\nu, r)$ VISUAL FREQUENCY WEIGHTS FOR THE 512×512 IMAGE *Lena*, CHANNEL *Y* AT 10 METERS.

Experiment 2: Correlation analysis between $\alpha(\nu, r)$ and $\hat{\alpha}(\nu, r)$. We employ the process shown in Fig. 7(a) for all the images of the CMU, CSIQ, and IVC Image Databases. In order to obtain $\hat{\alpha}(\nu, r)$, we measure the lineal correlation between the original $\alpha(\nu, r)$ applied during the F-ρSQ process and the recovered $\hat{\alpha}(\nu, r)$. Table 2 shows that there is a high similarity between the applied VFW and the recovered one, since their correlation is 0.9849, for gray-scale images, and 0.9840, for color images. In this section, we only expose the results for the CMU image database.

Table 2
CORRELATION BETWEEN $\alpha(\nu, r)$ AND $\hat{\alpha}(\nu, r)$ ACROSS CMU, CSIQ,
AND IVC IMAGE DATABASES.

Image Database	8 bpp gray-scale	24 bpp color
CMU	0.9840	0.9857
CSIQ	0.9857	0.9851
IVC	0.9840	0.9840
Overall	0.9849	0.9844

Fig. 8 depicts the PSNR difference (dB) of each color image of the CMU database, that is, the gain in dB of image quality after applying $\hat{\alpha}(\nu, r)$ at $d = 2000$ centimeters to the \hat{Q} images. On average, this gain is about 15 dB. Visual examples of these results are shown by Fig. 9, where the right images are the original images, central images are perceptual quantized images after applying $\alpha(\nu, r)$ and left images are recovered images after applying $\hat{\alpha}(\nu, r)$.

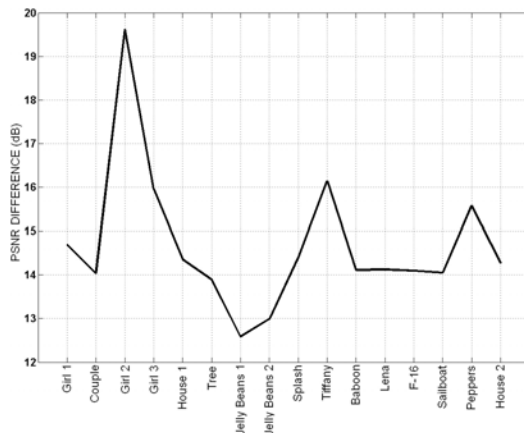
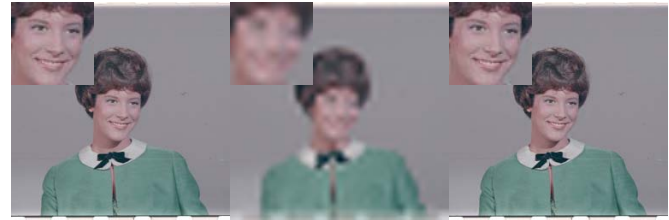


Fig. 8

PSNR DIFFERENCE BETWEEN \hat{Q} IMAGE AFTER APPLYING $\alpha(\nu, r)$ AND RECOVERED \hat{T} AFTER APPLYING $\hat{\alpha}(\nu, r)$ FOR EVERY COLOR IMAGE OF THE CMU DATABASE.

After applying $\hat{\alpha}(\nu, r)$, a visual inspection of these sixteen recovered images show a perceptually lossless quality. We perform the same experiment experiment for gray-scale and color images with $d = 20, 40, 60, 80, 100, 200, 400, 800, 1000$ and 2000 centimeters, in addition to test their objective and subjective image quality by means of the PSNR and MSSIM metrics, respectively.

In Figs. 10 and 11, green functions denoted as F- ρ SQ are the quality metrics of perceptual quantized images after applying $\alpha(\nu, r)$, while blue functions denoted as I- ρ SQ are the quality metrics of recovered images after applying $\hat{\alpha}(\nu, r)$. Thus,



(a) Girl 2



(b) Tiffany



(c) Peppers

Fig. 9

VISUAL EXAMPLES OF PERCEPTUAL QUANTIZATION. LEFT IMAGES ARE THE ORIGINAL IMAGES, CENTRAL IMAGES ARE FORWARD PERCEPTUAL QUANTIZED IMAGES (F- ρ SQ) AFTER APPLYING $\alpha(\nu, r)$ AT $d = 2000$ CENTIMETERS AND RIGHT IMAGES ARE RECOVERED I- ρ SQ IMAGES AFTER APPLYING $\hat{\alpha}(\nu, r)$.

either for gray-scale or color images, both PSNR and MSSIM estimations of the quantized image Q decrease regarding d , the longer d the greater the image quality decline. When the image decoder recovers \hat{Q} and it is perceptually inverse quantized, the quality barely varies and is close to perceptually lossless, no matter the distance.

5. Conclusions

In this work, we defined both forward (F- ρ SQ) and inverse (I- ρ SQ) perceptual quantizer using CBPF. We incorporated it to Hi-SET, testing a perceptual image compression system χ SET. In order to measure the effectiveness of the perceptual quantization, a performance analysis is done using thirteen assessments such as PSNR, MSSIM, VIF, WSNR or \mathcal{NR} PSNR, for instance, which measured the image quality between reconstructed and original images. The experimental results show that the solely usage of the Forward Perceptual Quantization improves the JPEG2000 compression and image

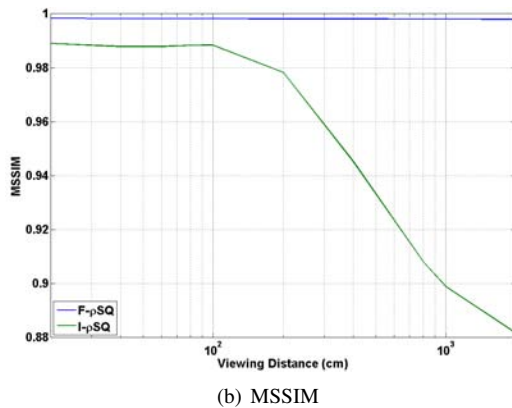
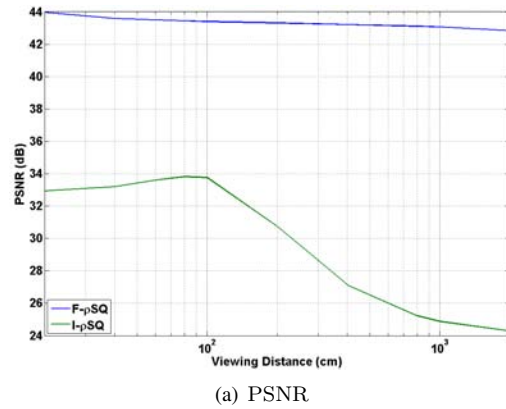


Fig. 10

PSNR AND MSSIM ASSESSMENTS OF COMPRESSION OF GRAY-SCALE IMAGES (Y CHANNEL) OF THE CMU IMAGE DATABASE. GREEN FUNCTIONS DENOTED AS $F\text{-}\rho\text{SQ}$ ARE THE QUALITY METRICS OF FORWARD PERCEPTUAL QUANTIZED IMAGES AFTER APPLYING $\alpha(\nu, r)$, WHILE BLUE FUNCTIONS DENOTED AS $I\text{-}\rho\text{SQ}$ ARE THE QUALITY METRICS OF RECOVERED IMAGES AFTER APPLYING $\hat{\alpha}(\nu, r)$.

perceptual quality. In addition, when both Forward and Inverse Quantization are applied into $Hi\text{-}SET$, it significantly improves the results regarding the JPEG2000 compression.

Acknowledgment

This work is supported by National Polytechnic Institute of Mexico by means of Project No. 20131312 granted by the Academic Secretary and the Committee of Operation and Promotion of Academic Activities (COFAA), National Council of Science and Technology of Mexico by means of Project No. 204151/2013, LABEX $\Sigma\text{-LIM}$ France, Coimbra Group Scholarship Programme granted by University of Poitiers and Region of Poitou-Charentes, France.

References

- [1] M. Boliek, C. Christopoulos, and E. Majani, *Information Technology: JPEG2000 Image Coding System*, JPEG 2000 Part I final committee draft version 1.0 ed., ISO/IEC JTC1/SC29 WG1, JPEG 2000, April 2000.

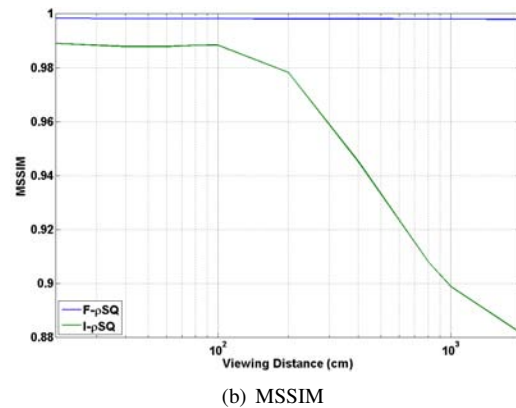
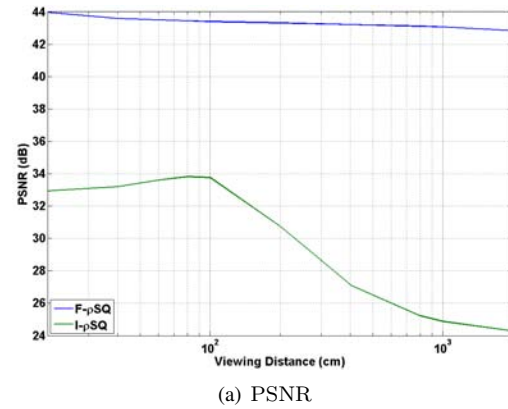


Fig. 11

PSNR AND MSSIM ASSESSMENTS OF COMPRESSION OF COLOR IMAGES OF THE CMU IMAGE DATABASE. GREEN FUNCTIONS DENOTED AS $F\text{-}\rho\text{SQ}$ ARE THE QUALITY METRICS OF FORWARD PERCEPTUAL QUANTIZED IMAGES AFTER APPLYING $\alpha(\nu, r)$, WHILE BLUE FUNCTIONS DENOTED AS $I\text{-}\rho\text{SQ}$ ARE THE QUALITY METRICS OF RECOVERED IMAGES AFTER APPLYING $\hat{\alpha}(\nu, r)$.

- [2] S. I. P. I. of the University of Southern California. (1997) The USC-SIPI image database. Signal and Image Processing Institute of the University of Southern California. [Online]. Available: <http://sipi.usc.edu/database/>
- [3] C. Research, École Polytechnique Fédérale de Lausanne, and Ericsson. (2001) JJ2000 implementation in Java. Cannon Research, École Polytechnique Fédérale de Lausanne and Ericsson. [Online]. Available: <http://jj2000.epfl.ch/>
- [4] T. Mitsa and K. Varkur, "Evaluation of contrast sensitivity functions for formulation of quality measures incorporated in halftoning algorithms," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 301–304, 1993.

SESSION
VIDEO PROCESSING, ANALYSIS AND
APPLICATIONS

Chair(s)

TBA

Trajectory scoring using JAABA in a noisy system

H. Chen†, B.Foley†, P. Marjoram

May 22, 2014

† *these authors contributed equally to this work.*

Corresponding author: P. Marjoram - pmarjora@usc.edu

Postal addresses: H. Chen and P. Marjoram: Dept. of Preventive Medicine, Keck School of Medicine, USC, Los Angeles, California 90089, USA. (HC: 071cht@gmail.com; PM: pmarjora@usc.edu)
B.Foley: Molecular and Computational Biology, Dept. of Biological Sciences, USC, Los Angeles, California 90089, USA. (brfoley76@gmail.com)

[This paper is intended for IPCV'14.]

Abstract

Many methods for generating tracking data for animals such as *Drosophila* assume idealized experimental conditions that are frequently difficult, expensive, or undesirable to replicate in the lab. In this paper we propose methods for improving robustness to non-ideal conditions. Our method involves an initial processing step in which tracks are constructed from noisy video data, followed by a subsequent application of machine learning algorithms to further improve performance. We demonstrate that our methods are capable of generating a substantial improvement in performance in fly identification, and allow for effective construction of tracks for individual flies. Furthermore, the methods we develop here represent a key first step in any subsequent attempt to automatically recognize interactions between flies, such as courtship and acts of aggression. As such, our algorithm provides a path for groups who wish to track fly, or characterize their behavior, in less than ideal conditions.

Keywords: tracking, machine learning, JAABA, behavior

Introduction

Behavioral studies commonly rely upon extensive time-series observation of animals, and characterization of their movement, activities and social interactions. Historically this involved scientists (or their students) recording observations by hand—a laborious and error-prone process. More recently, automation has promised to dramatically increase the quantity and detail of data collected, and a variety of methods have recently become popular in the important area of automated tracking, for example the CTRAX ethomics software [4], and the proprietary Ethovision [6].

Most available solutions demand restricted experimental conditions that may not be desirable for the question of interest, or feasible in the field, (or even the lab). For example, in *Drosophila melanogaster* experiments, it is common to restrict the possibility of flight, and use a backlit glass substrate for contrast [4]. A majority of *D. melanogaster* social interactions occur on food, and glass is not representative of their normal habitat. Additionally, many tracking algorithms perform poorly when the number of objects being tracked is not fixed. In such contexts it is difficult to determine whether a large “blob” of pixels in fact represents a single object or two overlapping objects. Such close contact happens commonly during aggression, courtship and mating events.

We are particularly interested in describing spontaneous group assembly. Here we consider data from a set of experiments in which we recorded fly behavior in an environment consisting of four food patches, modelled on a published experiment conducted with still cameras [9]. Each patch was recorded independently, and flies could freely move among patches, or be off patch (and thus not recorded). To model group assembly, we need to accurately count individuals on patches, and measure joining and leaving. We are currently able to detect objects (blobs, or putative flies) in video frames against a static background. This method is designed to be relatively robust to non-optimal experimental conditions.

Behavioral annotation requires that we move from static blobs, to individual-fly identification and tracking. We develop a two-stage process for this. First, we present an algorithm that enables us to assemble trajectories even through multi-fly blobs. We are then able to utilise these trajectories in freely available machine-learning behavioral annotation software. The Janelia Automatic Animal Behavior Annotator (JAABA) is a commonly used animal-behavior annotation software [5]. We use JAABA to manually flag errors in our tracking algorithm for “single fly” versus “multi-fly” blobs. This will enable trajectory correction and behavioural annotation.

Methods

Initial Video Processing: Videos are recorded using 4 high-resolution Grasshopper digital video cameras (Point Grey Research Inc., Richmond, Canada) simultaneously filming individual patches at 30hz, RGB, and 800×600 resolution. Videos are processed as single frames, by identifying blobs against an averaged background [2]. Blobs may contain from one to many individual flies, or be spurious artefacts. Features of the blobs are extracted using the cvBlobslib package [3]. The borders of the patch are defined manually, using the ImageJ software ellipse tool [1], and are calculated as length of the radius from centroid of the patch. All flies outside this radius are considered “off patch”. Lighting was ambient room lighting. Videos were recorded for one hour intervals, and a subset were scored for joining and leaving by hand, to evaluate accuracy.

Blobs are identified for each frame, or time T . The number of blobs, and blob statistics for each T , were output. Blob statistics include the blob X and Y centroids (B_X and B_Y); fitted-ellipse major and minor axes (B_A , B_B); and blob area (in pixels, B_P). Blobs with centroids outside the perimeter of the patch are excluded. Every blob is assigned a unique identifier within a frame (B_i). Each blob is subsequently assigned an inferred fly number (B_n , below).

Blobs and flies - Trajectory Assembly with Blob Uncertainty [TABU] : Flies are taken to be non-fissible blob units. We infer the number and identity of flies within blobs by tracking fusion and fission events. We construct tracks by making three simplifying assumptions (based on observation). First, flies do not often move a greater distance than their body length between frames. Second, the noise in blob area estimation is not large between consecutive frames (*i.e.*, less than half the actual fly area). Third, (on the scale of 30 frames a second) flies do not join and leave patches often—that is, we conservatively assume fly number does not change, unless there is no other explanation for the data. TABU is implemented in R [8].

Trajectories are constructed by initializing the first frame assuming each blob is a single fly. Subsequently we implement the following algorithm at each frame:

- 1. Identify Neighborhoods:** For each pair of frames T_t and T_{t+1} , we construct a graph by drawing edges between blobs that: a) are in different frames; and b) overlap. We define overlapping as having centroids within the distance of the major axis B_A of the blob ellipse. We define two degrees of overlapping: mutual and unidirectional. A mutual overlap occurs when the B_A of both blobs is longer than the distance between their centroids. If only one B_A is this long, the overlap is unidirectional. A “neighborhood” is defined as group of blobs linked by mutual edges.

2. Check “Joins” and “Leaves”: We test for probable joining and leaving events by examining blobs that are not in a neighborhood with other blobs, using the more-relaxed unidirectional overlap. Flies in blobs in T_t with no unidirectional matches in T_{t+1} are assumed to have left, and flies in blobs in T_{t+1} with no unidirectional matches in T_t are assumed to have joined. Otherwise, the unmatched blobs are assigned to their nearest unidirectional match.

3. Assign flies to blobs: In the simplest case, a neighborhood is comprised of a single blob in T_t and T_{t+1} . If so, all flies in the the blob at T_t are assigned to the blob at T_{t+1} . In more complex cases we assign flies between blobs to minimize the difference between summed fly-area and their respective B_p . Every fly inherits the blob-specific statistics of its assigned blob. During fission events if there are fewer flies than blobs we update fly numbers. Thus, we arrive at our count of flies. Each blob is also assigned a count of the number of flies it contains, B_N .

4. Update statistics: Each fly is assigned a number of fly-specific statistics. These include a unique index for each fly (F_j); fly area in pixels (F_p); and fly area from the fitted ellipse ($F_e = B_A B_B \pi$). Statistics are running averages, updated only when a fly is inferred to be in a single-fly blob. An error parameter, F_S , is also updated to alert us when there is a mismatch between observed blob properties and the inferred fly state—for instance, if the ratio between F_p and F_e is much smaller than 0.9, there is a high likelihood the blob contains multiple flies.

5. Resolve probable errors: For cases where error deviance F_S has grown too large, we attempt to reduce mismatch between imputed fly and blob matches by imputing leaving events, or evaluating group assignment.

We have found that this method gives us correct fly counts in blobs >85% of the time, but is subject to several systematic biases (see Results). For example it deals poorly with occlusion due to mounting which may last for seconds, and mating, which lasts for up to 20 minutes. It also may incorrectly infer several small flies instead of a single large fly. We therefore attempt a subsequent analysis aimed at correcting these remaining biases using Machine Learning [ML].

Machine Learning in JAABA and Trajectory Scoring

Once TABU has been applied, the trajectories become compatible with JAABA, allowing us to conveniently score behavior using its video annotation capabilities. We then fit ML algorithms. One (GentleBoost) is implemented within JAABA. The others (GradientBoost, logistic regression, and Support Vector Machine [SVM] with linear and Gaussian kernels [GSVM and LSVM]) we implemented ourselves using the Python Scikit-learn [7] package. For boosting we use decision stumps as the weak rules, and to ensure fair comparison default parameter values were used for all other methods.

Training of ML Algorithms: We used JAABA to calculate a number of internal single-frame fly statistics, as well as multi-frame window features. Window features are normalized to have mean 0 and variance 1. It is these features that are used for the ML classifiers. Users define behaviors, and score positive and negative cases for trajectories in the JAABA GUI, by observation in the video window. By treating blob state as a “behavior”, this allows us to train ML methods to model fly counts (and also to score the initial TABU blob calls).

Since the ML algorithms are binary classifiers, we score instances of “behavior” as a binary outcome “multifly”: multifly=1 for blobs labelled as having more than one fly; multifly=0 otherwise. We then fit ML classifiers using 3-fold cross-validation analysis in which the training data uses the manual annotations that we input using JAABA. After fitting, performance of each model is evaluated using accuracy, specificity, sensitivity, precision, and “area under curve”. Here, “accuracy” is defined as the proportion of times flies are are correctly identified as being in a multifly blob or not, for a total number of validation calls. All other performance measures follow the

usual statistical definitions. At the same time, we evaluate the performance of the TABU input trajectories by scoring whether our B_N statistics accurately describe blob fly count.

Results

We begin by evaluating the performance of the basic blob-recognition algorithm from [2], and the change in accuracy after processing the data with TABU, for the basic task of recognizing fly number and for joining and leaving events. The empirical 'real' results are obtained from manual annotation. Results are shown in Table 1. Let e be the estimated number of flies in a frame for a given method, n be the actual (manually annotated) number, and τ be the total number of frames. We estimate overall counting error, E , as $E = \frac{1}{\tau} \sum \frac{|e-n|}{n+1}$ (where the denominator is $n+1$ to avoid division by zero). This represents an approximate per-fly probability of being miscounted. Directionality, D , is calculated similarly, $D = \frac{1}{\tau} \sum \frac{e-n}{n+1}$, and demonstrates the chances of being consistently over- or under-counted. Joining or leaving events, 'Jump', are reported as the per-frame probability of either a change in blob number, or a trajectory starting or ending. Results are shown for three separate videos ('Rep').

Table 1: Performance of the blob algorithm output (Blob), and TABU trajectory output. Counting error (E), and Directionality (D) bias in counting is shown. Empirical (Real) and estimated fly patch-joining or leaving rates (Jump) are also shown for raw blob data and processed trajectories.

Rep	Blob E	TABU E	Blob D	TABU D	Real Jump	Blob Jump	TABU Jump
1	0.121	0.124	-0.074	0.100	0.012	0.127	0.047
2	0.143	0.093	-0.130	0.080	0.009	0.115	0.021
3	0.177	0.106	-0.150	0.080	0.013	0.085	0.037

By using TABU to create our trajectories, we have obtained more accurate, less biased, and less-noisy data (fewer spurious joining and leaving events) than the raw blob counts. Because of multi-fly blobs, the raw blob output tends to underestimate the true fly number when there is more than a single fly on a patch, while TABU has a slight bias towards over-counting. Even for the TABU output, there is a large excess of spurious joining and leaving events, offering potential for improvement through subsequent application of ML. The per-fly error rate (at least as reflected in per-fly count error) increases strongly for blob-counts (est=-0.06, df=107997, t=-206.3, $P < 0.001$). The change in error across fly number is much less pronounced (and the error rate approaches 0) in the TABU output (est=-0.007, df=107997, t=-22.2, $P < 0.001$) Figure 1.

We now investigate whether application of ML methods to our TABU trajectories can identify miscalled blob counts B_N . Three-fold cross-validation model-fit results are shown in Table 2. Here algorithms were trained using a period of 10K frames. We see that all models have an accuracy above 0.98. The two SVM models rank highly on almost all metrics, while logistic regression ranks poorly on most metrics. While JAABA is not top ranked on any metric, we note that it performs very well overall.

The critical practical question is whether models trained on one part of a video will be equally effective when applied to later periods of the same video, or to completely new video. Fly behavior is known to change over time, and varies among different genotypes and in different social contexts. We tested the performance of all algorithms on 3 videos that were not used in the training of the algorithm. This included different genotypes and sex ratios, as well as slightly different lighting and focus, than those the algorithms were trained on. Results are shown in Table 3. The performance of all ML methods dropped slightly under these new conditions. All the ML meth-

Figure 1: Heat map of the distribution of per-fly over- and under-counts as function of the number of flies on a patch for each frame across 3 test videos.

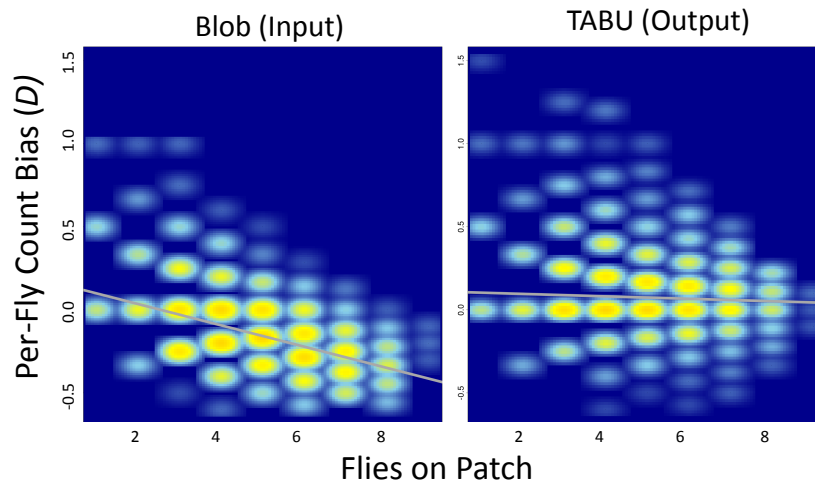


Table 2: Performance measures of ML algorithms on 3-fold cross validation. Ranks among ML methods for each performance score are given in brackets.

Algorithms	Accuracy	Specificity	Sensitivity	Precision	AUC
JAABA	0.994(2)	0.994(2)	0.994(3)	0.994(2)	-
GradientBoost	0.988(5)	0.987(3)	0.989(5)	0.987(5)	0.994(4)
Logistic	0.989(4)	0.984(5)	0.993(4)	0.985(3)	0.997(3)
ISVM	0.991(3)	0.985(4)	0.996(1)	0.986(4)	0.998(2)
gSVM	0.995(1)	0.995(1)	0.996(2)	0.995(1)	0.998(1)

ods improved upon the trajectory input data from TABU. The performance ranking of the ML algorithms remained broadly the same in this new data. The gSVM did very well, and logistic regression did relatively poorly. Again JAABA (GentleBoost) did very well overall.

Table 3: Evaluation of ML algorithm performance on non-training videos. Minimum to maximum range of scores shown for each.

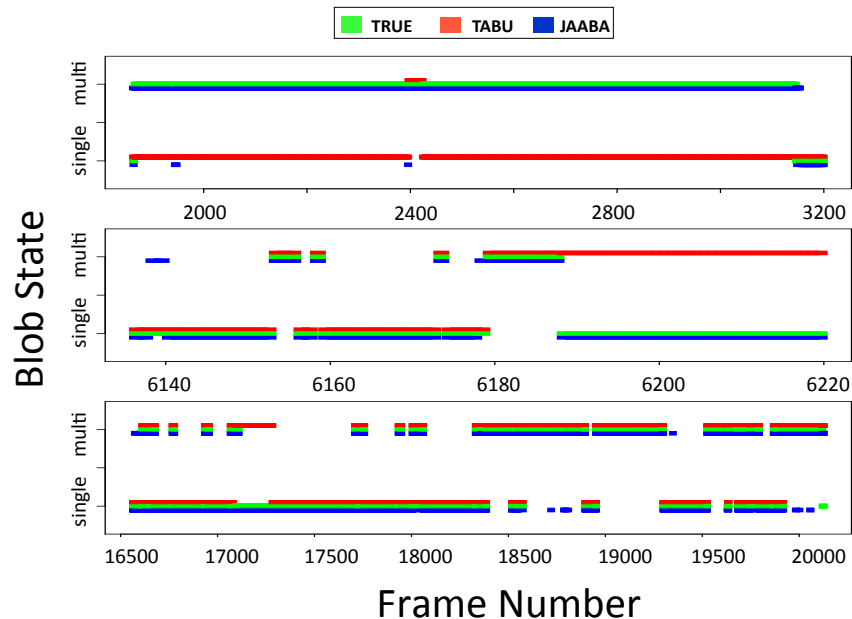
Algorithms	Accuracy	Specificity	Sensitivity
TABU	0.893 – 0.994	0.797 – 0.995	0.908 – 0.928
JAABA	0.987 – 0.994	0.988 – 0.996	0.936 – 0.999
GradientBoost	0.985 – 0.990	0.967 – 0.990	0.942 – 0.963
Logistic	0.986 – 0.991	0.972 – 0.996	0.890 – 0.993
ISVM	0.988 – 0.993	0.977 – 0.995	0.917 – 0.998
gSVM	0.987 – 0.994	0.980 – 0.997	0.908 – 0.999

Agreement of TABU with different learning algorithms over entire video

In order to better understand the sources of error, we investigated our trajectory data to find where TABU failed to identify transitions from either a single-fly blob to a multifly blob, or the reverse. Understanding when and why these errors occur will allow us to better finally correct our trajectories. In Figure 2, we show representative examples of sequences of consecutive frames in which TABU and JAABA have different predictions about blob state. In most cases both JAABA and

TABU agree, even for jumps between states as short as a single frame. In other cases when JAABA and TABU disagree, JAABA is more likely to be true.

Figure 2: Representative transitions from single-fly to multifly along trajectory fragments. The true state is shown in green, TABU predictions in red, and JAABA predictions in blue.



As shown in Figure 2, TABU's prediction errors mainly arise from the following two scenarios. If TABU fails to identify transitions from multifly to single fly, this misclassification will often last for a number of frames (*i.e.*, Figure 2, middle panel 6190-6220, bottom panel 17109-17282). Alternatively, when two flies come in patch together and rarely separate, TABU is more likely to classify this as a single fly rather than multifly blob. (Figure 2, top panel). In contrast, JAABA is relatively more sensitive to the correct blob state in both scenarios.

Discussion

In this paper we have developed a method for generating, and error correcting, tracking data from video recordings of *Drosophila* made in non-ideal conditions. Non-optimal conditions cause problems at the initial image processing stage, due to poor performance of background subtraction routines, occlusion caused by proximity between animals, and uncertainty in the number of objects that need to be tracked (*c.f.* [4]). This leads to subsequent poor performance of tracking data. However, imperfect conditions will apply for a majority of behavioral observation systems in nature. Even in many lab situations, experimenters often have to work with such conditions to collect relevant data. Our methods offer the potential for investigators to more successfully work with such data.

Our simple TABU tracking algorithm, by making a few realistic assumptions about the persistence of flies across frames and within blobs, greatly reduces the uncertainty of the initial image processing data from the algorithm of [2]. It allows us to count flies on patches with more certainty, and reduces the apparent degree of fly movement on and off of patches. Error rates are still non-zero, but it is clear that subsequent application of ML methods has the potential to increase correct allocation of flies among blobs from around 90% to over 98%.

Among the algorithms we evaluated, there is no clear winner among the ML methods in terms of performance. However, for ease of implementation, and robustly high performance, the Gentle-Boost algorithm natively implemented in JAABA represents a reasonable choice for future work. However, we note that use of JAABA requires fly tracking data as input, thereby necessitating pre-processing using an algorithm such as TABU before use. Such a pre-processing algorithm needs to be able to construct tracks successfully in non-ideal conditions, and when the number of objects being tracked is unknown, a problem that is known to be extremely challenging [4].

Our methods produce improved performance both in terms of accurate identification of the number of flies in a blob (and, therefore, the number of flies in a frame at any given moment), and in terms of generation of tracks for individual flies. Both of these types of information are crucial for analysis of fly (and other animal) group behavior. Flies are social animals, that actively aggregate and interact in groups. The sizes of these groups is therefore a key diagnostic of the behavior of those flies, and varies with factors such as genotype, sex ratio, etc. Therefore, the methods we present here provide the opportunity for researchers to use automated methods to generate large quantities of such data in an experimental context. A more difficult remaining challenge is to automatically recognize interactions between flies, such as courtship and acts of aggression. Methods (including JAABA) are being developed to attack this problem. Creating, and error correcting, fly trajectories is a necessary first step in taking advantage of this work.

Acknowledgements: The authors gratefully acknowledge funding from NSF and NIMH through awards DMS 1101060 and MH100879. The material contained in this paper reflects the views of the authors, and not necessarily those of NSF or NMH.

References

- [1] M.D. Abramoff, P.J. Magalhaes, and S.J. Ram. Image Processing with ImageJ. *Biophotonics International*, 11(7):36–42, 2004.
- [2] Reza Ardekani, Anurag Biyani, Justin E. Dalton, Julia B. Saltz, Michelle N. Arbeitman, John Tower, Sergey Nuzhdin, and Simon Tavaré. Three-dimensional tracking and behaviour monitoring of multiple fruit flies. *JOURNAL OF THE ROYAL SOCIETY INTERFACE*, 10(78), JAN 6 2013.
- [3] G. Bradski. *Dr. Dobb's Journal of Software Tools*.
- [4] Kristin Branson, Alice A. Robie, John Bender, Pietro Perona, and Michael H. Dickinson. High-throughput ethomics in large groups of *Drosophila*. *NATURE METHODS*, 6(6):451–U77, JUN 2009.
- [5] Mayank Kabra, Alice A. Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. JAABA: interactive machine learning for automatic annotation of animal behavior. *NATURE METHODS*, 10(1):64–U87, JAN 2013.
- [6] LPJJ Noldus, AJ Spink, and RAJ Tegelenbosch. EthoVision: A versatile video tracking system for automation of behavioral experiments. *BEHAVIOR RESEARCH METHODS INSTRUMENTS & COMPUTERS*, 33(3):398–414, AUG 2001.
- [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [8] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
- [9] J.B. Saltz and B.R. Foley. Natural genetic variation in social niche construction: social effects of aggression drive disruptive sexual selection in *Drosophila melanogaster*. *The American Naturalist*, 177:645–654, 2011.

Video Action Recognition Using an Optical Flow Based Representation

Samet Akpınar and Ferda Nur Alpaslan

Department of Computer Engineering, Middle East Technical University, Ankara, Turkey

Abstract – *In this study, a new model to the problem of video action recognition has been proposed. The model is based on temporal video representation for automatic annotation of videos. Video action recognition is a field of multimedia research enabling us to recognize the actions from a number of observations, where representation of temporal information becomes important. Visual, audio and textual features are important sources for representation. Although textual and audio features provide high level semantics, retrieval performance using these features highly depends on the availability and richness of the resources. Visual features such as edges, corners, interest points etc. are used for forming a more complicated feature, namely, optical flow. For developing methods to cope with video action recognition, we need temporally represented video information. For this reason, we propose a new temporal segment representation to formalize the video scenes as temporal information. The representation is fundamentally based on the optical flow vectors calculated for the frequently selected frames of the video scene. Weighted frame velocity concept is put forward for a whole video scene together with the set of optical flow vectors. The combined representation is used in the action based video segment classification. Proposed method is applied to significant data sets and the results are analyzed by comparing to the state-of-the art methods.*

Keywords: Video action recognition, content-based video information retrieval, optical flow, temporal video segment representation, weighted frame velocity

1 Introduction

Video action recognition is a field of multimedia research enabling us to recognize the actions from a number of observations. The observations on video frames depend on the video features derived from different sources. While textual features include high level semantic information, they cannot be automated. The recognition strongly depends on the textual sources which are commonly created manually. On the other hand, audio features are restricted to a supervisor role. As the audio does not contain strong information showing the actions conceptually, it can be used as an additional resource supporting visual and textual information. Visual video features provide the basic information for the video events or actions. Although it is difficult to obtain high levels of semantics by using visual

information, a convincing way to construct an independent fully automated video annotation or action recognition model is to utilize visual information as the central resource. This way takes us to content-based video information retrieval.

Content-based video information retrieval is the automatic annotation and retrieval of conceptual video items such as objects, actions, events, etc. using the visual content obtained from video frames. There are various methods to extract visual features and use them for different purposes. The visual feature sets they use vary from static image features (pixel values, colour histograms, edge histograms, etc.) to temporal visual features (interest point flows, shape descriptors, motion descriptors, etc.). Temporal visual features combine the visual image features with the time information. Representing video information using temporal visual features generically means modelling the visual video information with temporal dimension. i.e., constructing temporal video information.

We need to represent the temporal video information formally for developing video action recognition methods. Visual features such as corners, visual interest points etc. of video frames are the basics for constructing our model. These features are used for constructing a more complicated motion feature, namely, optical flow. In our work, we propose a new temporal video segment representation method to retrieve video actions for formalizing the video scenes as temporal information. The representation is fundamentally based on the optical flow vectors calculated for the frequently selected frames of the video scene. Weighted frame velocity concept is put forward for a whole video scene together with the set of optical flow vectors. The combined representation is used in the action based temporal video segment classification. The result of the classification represents the recognized actions.

The main contribution of this work is the proposed temporal video segment representation method. It is aimed to be a generic model for temporal video segment classification for action recognition. The representation is based on optical flow concept. It uses the common way of partitioning optical flow vectors according to their angular features. An angular grouping of optical flow vectors is used for each selected frame of the video. We propose the novel concept of *Weighted Frame Velocity* as the velocity of the cumulative angular grouping of a temporal video segment in order to represent the motion of the frames of the segments more descriptively.

The outline of the paper is as follows. In Chapter 2, related work is proposed. Chapter 3 discusses the temporal

segment representation. Optical flow is described in Chapter 4 and optical flow based segment representation is discussed in Chapter 5. In Chapter 6, experiments and results are presented. Finally, in Chapter 7, the conclusion is proposed.

2 Related Work

There are different approaches followed for the representation of temporal video segments for content-based video information retrieval problems such as video action recognition, event detection, cut detection, etc. The studies in [10, 11, 12, 13] focus on the perception of the visual world and bring us facts about how to detect the visual features and in which context more philosophically. Regarding the visual features, mentioned approaches can generally be figured out. Key-frame, bag-of-words, interest points and motion based approaches are the groups of approaches reflecting the way of representation.

Key-frame based representation approaches focus on detecting key frames in the video segments in order to use them in classification. This kind of representation is used in [1, 2, 3, 14] for video scene detection and video summarization. The study of [1] contains the segmentation of videos into shots and key-frames are extracted from these shots. In order to overcome the difficulty of having prior knowledge of the scene duration, the shots are assigned to visual similarity groups. Then, each shot is labelled according to its group and a sequence alignment algorithm is applied for recognizing the shot labels change patterns. Shot similarity is estimated using visual features and shot orders are kept while applying sequence alignment. In [2], a novel method for automatic annotation of images and videos is presented with keywords among the words representing concepts or objects needed in content-based image retrieval. Key-frame based approach is used for videos. Images are represented as the vectors of feature vectors containing visual features such as colour, edge, etc. They are modelled by a hidden Markov model, whose states represent concepts. Model parameters are estimated from a training set. The study proposed in [3] deals with automatic annotation and retrieval for videos using key frames. They propose a new approach automatically annotating video shots with semantic concepts. Then, the retrieval carried out by textual queries. An efficient method extracting Semantic Candidate Set (SCS) of video shots is presented based on key-frames. Extraction uses visual features. In [14], an innovative algorithm for key frame extraction is proposed. The method is used for video summarization. Metrics are proposed for measuring the quality.

Histogram based bag-of-words (BoW) approaches represent the frames of the video segments over a vocabulary of visual features. [4, 5] are the examples of such approaches. [4] proposes a method interpreting temporal information with the BoW approach. Video events are conceptualized as vectors composed of histograms of visual features, extracted from the video frames using BoW model. The vectors, in fact, can be behaved as the sequences, like strings, in which histograms are considered as characters. Classification of these sequences having difference in length, depending on the video scene length, is carried out by using SVM classifiers with a string kernel that uses the Needleman-Wunsch edit distance. In [5], a

new motion feature is proposed, Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW) implementing motion relativity and visual relatedness needed in event detection. Concerning the ERMH-BoW feature, relative motion histograms are formed between visual words representing the object activities and events.

Despite their performance issues in terms of time, above approaches lack the flow features and temporal semantics of motion although they are efficient in spatial level. On the other hand, motion based approaches deal with motion features which are important in terms of their strong information content and stability over spatio-temporal visual changes. Motion features such as interest points, optical flow, etc. are used for modelling temporal video segments. [6, 7, 15, 35] are the studies using motion features. [6] proposes a new framework in order to group the similar shots into one scene. Motion characterization and background segmentation are the most important concepts in this study. Motion characterization results in the video representation formalism while background segmentation provides the background reconstruction which is integrated to scene change detection. These two concepts and the colour histogram intersection together become the fundamental approach for calculating the similarity of scenes. The study of [7], presents a new approach which implements motion estimation in video scenes. The representation of video motion is carried out by using some sort of particles. Each particle is an image point with its trajectory and other features. In order to optimize the particle trajectories, appearance stability along the particle trajectories and distortion between the particles are measured. The motion representation can be used in many areas. It cannot be constructed using the standard methods such as optical flow or feature tracking. Optical flow is a spatio-temporal motion feature describing the motion of visual features. Optical flow based representation is especially strong for video segment classification. [15, 35] present methods for representing video segments with optical flow. [15] proposes a representation structure based on direction histograms of optical flow. In [35], video segments are tried to be represented by using histogram of oriented optical flow (HOOOF). By the help of this representation, human actions are recognized by classifying HOOOF time-series. For this purpose, a generalization of the Binet-Cauchy kernels to nonlinear dynamical systems (NLDS) is proposed.

Temporal video segment classification is an important sub problem in content-based video information retrieval addressing video action classification in our study. By definition, it is the classification of scenes in a video. The classification highly depends on the representation of temporal video information and the classification methods working on this representation.

[33, 34, 37, 39] propose the approaches based on 3D interest points. These methods tackle the problem of video segment classification by putting new interest points or visual features forward by enriching with time dimension. Therefore, the features in the studies can be conceptualized as space-time shapes.

The methods proposed in [8, 30] views the problem from the point of spatio-temporal words. The segments are

seen as bag-of-features and make the classification according to the code words.

[15, 35, 38, 40, 41] present optical flow based methods for video segment classification. Optical flow histograms are constructed and utilized in segment representation. By using this representation segment classification is carried out.

3 Temporal Segment Representation

Temporal video segment representation is the problem of representing video scenes as temporal video segments. While this problem generally runs through the video information including visual, audio and textual features, our study deals with visual features only. Mentioned problem is originated from representing the temporal information. Temporal information provides a combined meaning composed of time and magnitude for a logical or physical entity. Robot sensor data, web logs, weather, video motion and network flows are common examples of temporal information. Independent from domain, both representation and processing methods of temporal information is important in the resulting models. Regarding the processing methods, prediction, classification and mining can be considered as first comers for the temporal information. In most cases, the representation is also a part of the processing methods due to the specific problem. While the representation and processing methods are handled together, the focus is especially on the processing methods rather than the representation in these cases. Temporal data mining and time series classification can be exemplified for the approaches on temporal information retrieval.

The types of the features and their quality on describing the domain knowledge also influence the temporal information processing and its application. Also, having high dimensionality makes the effective representation of temporal information with more complicated features important. Therefore, feature definitions, construction and feature extraction methods play an important role in processing the temporal information. As the focus here is feature extraction and construction, the improvements are measured with common methods.

In content-based video information retrieval, visual video data behaves like temporal information containing frame sequences over time. Each frame of the video has its visual information along with its time value. The temporal information representation highly depends on the visual content of video frames. The basic and the most primitive representation of temporal video information can be done by using the video with all pixel intensities of all frames. While this representation includes the richest visual information, processing and interpreting information is impractical. In a 600x480 frame size for a 10 seconds scene (30 fps), 86.4M features exist with this approach. Therefore, there is a need for efficient representation formalisms.

Key-frame based representation is one of the candidate approaches for representing temporal information in videos. For each scene, a key-frame is selected based on some calculations using visual features. The entire scene is represented and feature size of the representation is decreased by using this key frame. But, there is an important

problem in key-frame based approaches; lack of the important information resulting from the motion in videos.

Another approach is bag-of-words approach for frame sequences. In this kind of representation, frames are behaved as code words obtained from grouping of the frames according to the visual features. With these code words, frame sequences are represented as sentences. This kind of representation contains temporal nature of the scenes. But, the most important disadvantage of this representation is the restricted nature of code words. Representing a visually rich frame with a label means losing an important amount of information. The representation is restricted with the variety of the code words. Therefore, limitless types of frames will be reduced to very limited number of labels.

Interest points based representation is an alternative formalism for temporal video information. Interest points are the "important" features that may best represent the video frames invariant from the scale and noise. This representation alternative is very successful in reducing the huge frame information into small but descriptive patterns. But, it is again disadvantageous in detecting motion features despite its descriptiveness. As the motion features include flow with time, it is important to track the features along the time. Using interest points for representation lacks the motion based information.

State-space methods are also used for representing temporal video information. The state-space methods define features which span the time. The space-time interest point concept is proposed by [34]. Interest points which are spatially defined and extracted in 2D are extended with time. With this extension, interest points gain a 3D structure with time. Therefore, a space-time 3D sketch of frame patterns can be obtained and they are ready for processing. State-space approaches best fit the representation of video information temporally as they can associate the time with the visual information in a descriptive and integrated way.

In our study, a state-space based representation approach is proposed. Optical flow is the motion feature - integrating time with visual features - utilized for constituting the state-space method.

4 Optical Flow

Theoretically, optical flow is the motion of visual features such as points, objects, shapes etc. through a continuous view of the environment. It represents the motion of the environment relative to an observer. James Jerome Gibson firstly introduced the optical flow concept in 1940s, during World War II [16]. He was working on pilot selection, training, and testing. He intended to train the perception of pilots during the war. Perception was considered for the effect of the motion on the observer. In this context, shape of objects, movement of entities, etc. are handled for perception. During his study on aviation, he discovered optical flow patterns. He found that the environment observed by the pilot tends to move away from the landing point, while the landing point does not move according to the pilot. Therefore, he joined this concept with the pilot perception on the observed environment. In Figure

1, landing plane is shown with optical flow departing from the landing point using the pilot view.

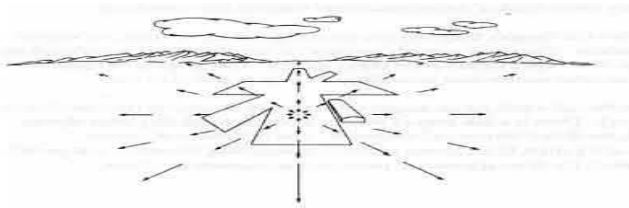


Figure 1. Landing plane with optical flow

In the perception of an observer, there may be two options for approaching/departing optical flow around a point. In the first option, the observer may be moving through the target point. This makes the optical flow departing from the point. In the second option, the environment around the point may be moving through the motionless observer. This also gives the same effect, having the optical flow departing from the point. These two options are also valid for approaching optical flow. If the observer departs from the target point or the point departs from the motionless observer, the optical flow is seen as approaching through the point.

In video domain, optical flow is commonly known as the apparent motion of brightness patterns in the images. More specifically, it can be conceptualized as the motion of visual features such as corners, edges, ridges, textures etc. through the consecutive frames of a video scene. Optical flow, here, is materialized by optical flow vectors. An optical flow vector is defined for a point (pixel) of a video frame. In optical flow estimation of a video frame, selection of “descriptive” points is important. This selection is done using visual features. It is clear that using an edge point or corner point is more informative than using an ordinary point semantically as the motion perception of human is based on prominent entities instead of ordinary ones. Optical flow vectors are, then, the optical flow of video frame feature instances instead of all frame points.

Two problems arise in the optical flow estimation of video frames: (1) detection and extraction of the features to be tracked, (2) calculation of the optical flow vectors of the extracted features. Optical flow estimation aims to find effective solutions to these problems. Calculation of optical flow vectors of the extracted features can be reduced to the following problem; “Given a set of points in a video frame, finding the same points in another frame”

4.1 Derivation of Optical Flow

There are various approaches concerning the estimation of optical flow. *Differential, region-based, energy-based, and phased-based* methods are the main groups of approaches [17]. All of these groups include many algorithms proposed so far. Each of these algorithms reflects the theoretical background of its group of approach.

Here, the meaning of optical flow estimation is discussed from a *differential* point of view. The explanation is based on the change of pixels with respect to time. The solution of the problem can be reduced to the solution of the following equation [19]:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) \quad (1)$$

The equation is written for a point in a video frame. The point is assumed to change its pixel value over time. (x, y, t) is defined as the 3D point composed of the 2D coordinates (pixel value) of the given point at time t . I represents the intensity function giving the image intensity value of a given pixel value at a given time. The equation is based on the assumption that enormously small amount of change on the pixel position of the point in enormously small amount of time period converges to zero change in the intensity value. In other words, the intensity value of a pixel in a frame is equal to the intensity value of another pixel having the same point (the point in the former pixel) in the next frame. The point moves enormously small amount of distance (pixel change) in enormously small amount of time.

LHS of the equation is expanded by using the Taylor Series Expansion [19]. By some further calculations, we obtain the following equations (2,3). The variables represented by I_x, I_y, I_t are the derivatives of intensity function according to all dimensions.

$$I_x V_x + I_y V_y = -I_t \quad (2)$$

$$\nabla I \cdot \vec{V} = -I_t \quad (3)$$

Now, the problem converges to the solution of \vec{V} . The solution will be the estimation of optical flow. As there are two unknowns in the equation, it cannot be solved; additional constraints and approaches are needed for solution. This problem is known as *aperture problem*.

4.2 Algorithms

Many algorithms according to different approaches have been proposed for optical flow estimation. According to [17], optical flow estimation algorithms can be grouped according to the theoretical approach while interpreting optical flow. These are *differential techniques, region-based matching, energy-based methods and phase-based techniques*.

Differential Techniques:

Differential techniques utilize a kind of velocity estimation from spatial and temporal derivatives of image intensity [17]. They are based on the theoretical approach proposed by [19]. The proposed approach results in the equation (3). Differential techniques are used for solving the problem generally represented by this equation. Horn-Shunck method [19] is a fundamental method among the differential techniques. Global smoothness concept is also used in the approach. Lucas-Kanade method [24] is also an essential method solving the mentioned differential equation for a set of neighboring pixels together by using a weighted window. [25, 29] use second order derivatives generating the optical flow equations. Global smoothness concept is also used as well as the Horn-Shunck method. [18] proposes a distance based method efficient for real-time systems. The method is analyzed according to time-space complexity and its tradeoff. [20] suggests a classical differential approach. But, it is combined with correlation based motion descriptors.

Region-Based Matching:

Region-based matching approaches alternate the differential techniques in case differentiation and numerical operations is not useful due to noise or small number of frames [17]. In region-based matching, the concepts such as velocity, similarity, etc. are defined between image regions. [21, 27] propose region-based matching methods for optical flow estimation. In [21], the matching is based on Laplacian pyramid while [27] recommends a method based on sum of squared distance computation.

Energy-Based Methods:

Energy-based methods are based on the output energy of filters tuned by the velocity [17]. [26] proposes an energy-based method fitting spatiotemporal energy to a plane in frequency space. Gabor filtering is used in the energy calculations.

Phase-Based Techniques:

Different from energy-based methods velocity is defined as filter outputs having phase behavior. [28, 23, 22] are the examples of phase-based techniques using spatiotemporal filters.

5 Optical Flow Based Segment Representation

In this study, an optical flow based temporal video information representation is proposed. Optical flow vectors are needed to be calculated for the selected sequential frames. Optical flow estimation is important as the basic element of the model is optical flow vectors. As mentioned in Chapter 4, detection of features and estimation of optical flow according to these features are the main steps of optical flow estimation. The methods and approaches for both steps are discussed below.

5.1 Optical Flow Estimation

In our approach, *Shi-Tomasi* algorithm proposed in [32] is used for feature detection. As it is mentioned before, Shi-Tomasi algorithm is based on Harris corner detector [31] and finds corners as interest points. Shi-Tomasi algorithm uses the eigenvalues of the Harris matrix. In this context, it differs from Harris corner detector. The algorithm assumes that minimum of two eigenvalues of Harris matrix determines the cornerness (C) of the point. Therefore, the corner decision is done using the eigenvalues of the matrix. Shi-Tomasi algorithm gives more accurate results compared with Harris detector. The algorithm is also more stable for tracking. For estimating optical flow, Lucas-Kanade algorithm is selected [24]. With videos having sufficient information and excluding noise, Lucas-Kanade algorithm is successful. The algorithm works for the corners obtained from Shi-Tomasi algorithm. Basically, the following function should be minimized for each detected corner point as seen in differential approaches:

$$\epsilon(\delta x, \delta y) = I(x, y) - I(x + \delta x, y + \delta y) \quad (4)$$

With suitable δx and δy , optical flow vectors can be obtained. But, aperture problem is not solved yet with this

minimization. The solution approach for aperture problem is reflected to the function definition $\epsilon(\delta x, \delta y)$ as follows.

$$\epsilon(\delta x, \delta y) = \sum_{u_y-w_y}^{u_y+w_y} \sum_{u_x-w_x}^{u_x+w_x} [I(x, y) - I(x + \delta x, y + \delta y)] \quad (5)$$

Summation on $x - y$ direction is a solution for the aperture problem. By using a window w centering the point (x, y) , the estimation of optical flow of the point is extended with the neighboring points.

In our approach, Lucas-Kanade algorithm is applied to the corner points detected with Shi-Tomasi algorithm.

Video frames are selected according to a frequency of 6 frames/sec (30 fps videos are used) from "Hollywood Human Actions" dataset [9]. In Figure 2, two frequently sequential frames obtained from the mentioned dataset are shown.



Figure 2: Consecutive frames for optical flow estimation

Figure 3 shows the optical flow vectors estimated for the detected points in the former video frame in the sequence.



Figure 3: Frame with optical flow vectors

In our method, optical flow vectors are calculated for every detected point in all frequently selected frames. The set of optical flow vectors is the temporal information source for our representation.

The model below forms the back bone for our representation formalism. Optical flow vector set with an operator constructs the representation.

$$R = [S(V), \Phi] \quad (6)$$

$S(V)$ is the set of optical flow vectors while Φ is the descriptor operator. Operator defines the relation of the elements of the optical flow vector set of the frames. This relation exposes the temporal representation of video information. The operator may change according to the complexity of the model. It may vary from just counting the

vectors to complex relations between the optical flow vectors. This generic representation can easily be adapted to different problems such as segment classification or cut detection. Choice of the operator and the optical flow representation may change drastically in different problems.

5.2 Proposed Representation

Usage of optical flow in video information representation is encountered in many studies including [36, 37, 35]. These studies are the state-of-the-art techniques motivating us for an optical flow based representation. Optical flow histogram is the most common way of optical flow based video representation. In [36], optical flow histograms are used for characterizing the motion of a soccer player in a soccer video. A motion descriptor based on optical flow is proposed and a similarity measure for this descriptor is described. The study of [37] uses optical flow by splitting it into horizontal and vertical channels. The histogram is calculated on these channels. Each channel is integrated over the angularly divided bins of optical flow vectors. In [35], histogram of oriented optical flow (HOOF) is simply used according to angular segments for each frame. The feature vectors are constructed with these angular values and combined for all frames of the video segment. The essential part for contribution here is the classification method. The classification is done with a proposed novel time-series classification method including a metric for comparing optical flow histograms. The study in [38] proposes an optical flow based representation which groups the optical flow vectors of whole video segment according to angular values. Then, average histogram is computed for each of these angular groups. The resulting histogram is the feature vector.

In our approach, histogram based optical flow approaches are enriched with a newly defined velocity concept, *Weighted Frame Velocity*. The idea, here, is originated from the inadequacy of optical flow histograms for interpreting information. Using optical flow histogram is discarded as the most important drawback of using histograms in segment representation is that the histogram similarity does not always mean the real similarity for motion characterization. Optical flow vectors are divided into angular groups and according to these groups, optical flow vectors are summed and integrated with the new velocity concept instead of a histogram based approach.

Estimating the optical flow vectors for each frame is the first step. Then, the equation (6) giving the generic representation is adapted to segment representation. In this aspect, Φ is the operator defining the relations between the optical flow vectors and giving their meaning for representing the video segment composed of the set of optical flow vectors $S(V)$.

In our adaptation of the above representation to segment representation, the description of Φ is important. The parameters used in the definition are shown in Table 1.

Table 1. Segment representation model parameters

PARAMETER	DEFINITION
F	Set of frames in the video segment
$S(V_f)$	Set of optical flow vectors in frame f
$S(V_f, \alpha, \beta)$	Set of optical flow vectors having angle between $\alpha - \beta$ in frame f
$A(\alpha, \beta)$	Weighted frame velocity of the whole segment direction having angle between $\alpha - \beta$
$\tau_f(\alpha, \beta)$	Threshold function for optical flow vectors having angle between $\alpha - \beta$ in frame f
$V(r, \angle\varphi)$	Optical flow vector having magnitude r and angle φ

The parameters above are the basic building blocks for constructing the representation model and the descriptor operator Φ . The following definitions are done for this purpose. The definition of $S(V_f, \alpha, \beta)$ is made as follows:

$$S(V_f, \alpha, \beta) = \{V(r, \angle\varphi) \in V_f \mid \alpha < \varphi \leq \beta\} \quad (7)$$

Let's assume that $|F| = n$, m is the number of angle intervals and l is the length of the video segment in terms of seconds. With these assumptions, the representation of a video segment using average of optical flow vectors with angular grouping can be formulized as:

$$R = [\parallel \sum_{S(V_{vf, \alpha_1, \alpha_2})} V(r, \angle\varphi) \parallel, \parallel \sum_{S(V_{vf, \alpha_2, \alpha_3})} V(r, \angle\varphi) \parallel, \dots, \parallel \sum_{S(V_{vf, \alpha_m, \alpha_{m+1}})} V(r, \angle\varphi) \parallel] \quad (8)$$

Above vector representation is composed of m dimensions each of which is the magnitude of the sum of optical flow vectors for angle intervals. This is the common way of optical flow representation except the usage of vectors instead of histograms. This representation is descriptive as it utilizes the movement of a segment in different angel intervals by using the vector sum and magnitude calculation. But, it lacks the temporal information in terms of velocity. This means that the flow details throughout the frame sequence are discarded by only looking at the resulting direction and magnitude information. If this vector is extracted for each frame and combined for solving the problem as it is done in [35], curse of dimensionality problem arises. The dimension of the

resulting vector will be $m \times n \times l$. For a 30 fps video of 5 seconds length with 30 angular intervals, for example, a vector of 4500 features is obtained for representing a segment. Using a frequency filter of 0.2 (6 frames selected from a second of the video) will decrease the dimension into 900, but the problem will not be able to be solved yet. This yields to the need for tackling the curse of dimensionality problem as it is handled in [35] with the newly proposed time series classification method including the new distance metric for the feature vectors.

In our approach, we enrich the representation to make the temporal information more descriptive without causing the curse of dimensionality problem. For this purpose, a new component is needed for the above feature representation based on movement magnitude of the segment in different directions. Velocity is selected as the fundamental idea for the new component as the velocity of the frames strongly affects the nature of video motion such as in walk and run events. For this purpose, weighted frame velocity concept is proposed. Abstractly, the velocity component is added to the feature vector to contain distance-velocity pair.

Weighted frame velocity is a metric which measures the velocity of a segment in a given dimension. It is weighted with the vector count in its direction. Theoretically, weighted frame velocity is formulated inspiring from the general velocity calculation $V = \frac{\Delta d}{\Delta t}$:

$$A(\alpha, \beta) = \frac{\sum_{i=0}^{n-1} [\| \sum_{S(V_{f_i, \alpha, \beta})} V(r, \angle \varphi) \| \cdot |S(V_{f_i, \alpha, \beta})|]}{\sum_{i=0}^{n-1} |S(V_{f_i, \alpha, \beta})|} \quad (9)$$

The equation (9) calculates the weighted distance for each angular interval of each frame. Weight concept, here, is the weight of the frame to the segment. The weighted distances are summed up and averaged according to the number of vectors in the segment. The resulting value is the weighted velocity of the frames.

When this approach is analyzed, one can notice that another problem occurs. As the velocity is weighted according to the number of vectors in the given angle interval of the frame, the noise or errors resulting from optical flow estimation and insignificant number of vectors in one dimension unfairly dominate the values of that feature. In order to avoid this problem, thresholding is used as a common way of noise reduction. Therefore, a threshold function depending on the frame and angle interval is proposed to be used in the weighted frame velocity function.

$$\tau_{f_i}(\alpha, \beta) = \begin{cases} \frac{S(V_{f_i, \alpha, \beta})}{S(V_{f_i, 0, 2\pi})}, & \frac{S(V_{f_i, \alpha, \beta})}{S(V_{f_i, 0, 2\pi})} < C \\ 1, & otherwise \end{cases} \quad (10)$$

The above function is based on the ratio of the optical flow vectors of the given angle interval for the given frame. This ratio's being smaller or bigger according to the threshold value C directly determines the result of the

function. At this point, estimation of threshold becomes important. The estimation will be done during the classification phase. Thus, the weighted frame velocity function is updated accordingly.

$$A(\alpha, \beta) = \frac{\sum_{i=0}^{n-1} [\| \sum_{S(V_{f_i, \alpha, \beta})} V(r, \angle \varphi) \| \cdot |S(V_{f_i, \alpha, \beta})| \cdot \tau_{f_i}(\alpha, \beta)]}{\sum_{i=0}^{n-1} |S(V_{f_i, \alpha, \beta})|} \quad (11)$$

The function affects the weighted contribution of each frame into the velocity of the segment in an angle interval according to whether its vectors' are noisy or not.

As it is mentioned before, the weighted frame velocity is, now, a new component of the feature vector representation based on the movement of the segment. Thus, the new representation is as follows:

$$R = [A(\alpha_1, \alpha_2), \| \sum_{S(V_{\varphi f, \alpha_1, \alpha_2})} V(r, \angle \varphi) \|, A(\alpha_2, \alpha_3), \| \sum_{S(V_{\varphi f, \alpha_2, \alpha_3})} V(r, \angle \varphi) \|, \dots, A(\alpha_m, \alpha_{m+1}), \| \sum_{S(V_{\varphi f, \alpha_m, \alpha_{m+1})} V(r, \angle \varphi) \|] \quad (12)$$

Now, the operator Φ in the generic optical based representation model $R = [S(V), \Phi]$ is defined in this specific problem. The operator maps the optical flow vector set $S(V)$ to the feature vector R for a video scene.

$$\Phi: S(V) \rightarrow R \quad (13)$$

The function of the above mapping is shown in the obtained final representation. Mainly, it constructs the representation by applying the operator to the optical flow vectors. The operator Φ , in fact, is the symbolic representation of our method.

The practical use of the representation is classifying the segments. The representation is used for each video segment and has the size $m \times 2$. The segment classification, constant estimations, and experiments with results and comparisons will be held in Chapter 6.

6 Experiments and Results

Temporal segment classification for action recognition uses the vector representation proposed in Chapter 5. Support Vector Machines (SVM) is used for non-linear classification. Gaussian radial basis function - using standard deviation σ for two feature vectors x_i, x_j - is selected as SVM kernel.

$$K(x_i, x_j) = e^{(-\frac{1}{2\sigma^2} \|x_i - x_j\|^2)} \quad (14)$$

Hollywood Human Actions dataset [9] is used for evaluation. Hollywood dataset includes video segments composed of human actions from 32 movies. Each segment

is labeled with one or more of 8 action classes: “AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, and StandUp”. While, the test set is obtained from 20 movies, training set is obtained from 12 other movies different from those in the test set. The training set contains 219 video segments and the test set contains 211 samples with manually created labels.

After the optical flows are estimated, the calculations for constructing feature vectors are carried out accordingly and feature vectors are obtained for the test data. The number of angular intervals is taken as 30 as in [35]. The threshold C in the threshold function below, as discussed in Chapter, was determined experimentally.

$$\tau_{f_i}(\alpha, \beta) = \begin{cases} \frac{S(V_{f_i}, \alpha, \beta)}{S(V_{f_i}, 0, 2\pi)}, & \frac{S(V_{f_i}, \alpha, \beta)}{S(V_{f_i}, 0, 2\pi)} < C \\ 1, & \text{otherwise} \end{cases} \quad (15)$$

The result of the experiments for determining the best threshold value which is 0.025 is shown in Figure 4. The experiments on Hollywood data set were carried out using this threshold value.

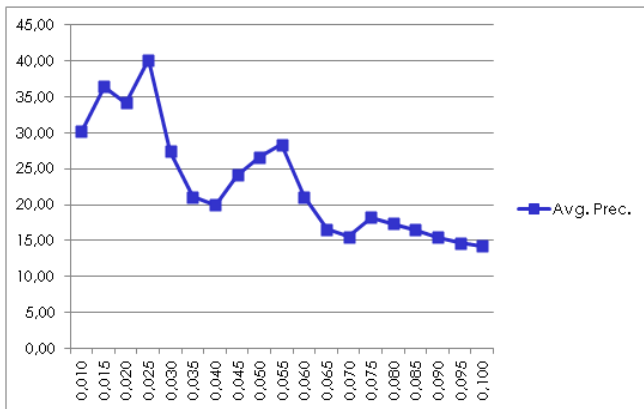


Figure 4. Threshold estimation in segment representation for Hollywood Human Actions

The results are compared to the ones obtained by the study in [9] which uses a state-of-the-art method and can be a substantial reference for video segment classification based on space-time points, as shown in Table 2.

Table 2. Comparison of the results of video segment classification for Hollywood Human Actions

EVENTS	ACCURACY	PRECISION	
	This Paper	Laptev et al. 2008	This Paper
StandUp	76.7%	50.5%	40.0%
SitDown	85.7%	38.6%	42.9%
HandShake	90.9%	32.3%	40.0%
Hug	89.5%	40.6%	44.4%
SitUp	86.2 %	18.2%	33.3%

Accuracy values are high, as the ratio of each event is low in total, as shown in the table. Except the standup event, better precision results are obtained. Especially, concerning “SitUp” action, the success rate is doubled.

Another comparison is made with the popular state-of-the-art Weizmann data set. The data set contains the actions “walk”, “run”, “jump”, “side”, “bend”, “one-hand wave”, “two-hands wave”, “pjump”, “jack”, and “skip”.

First, the threshold estimation is carried out for this set again. As shown in Figure 5, the threshold values 0.020 – 0.025 gave the best results. These values were used in evaluating the results over this data set.

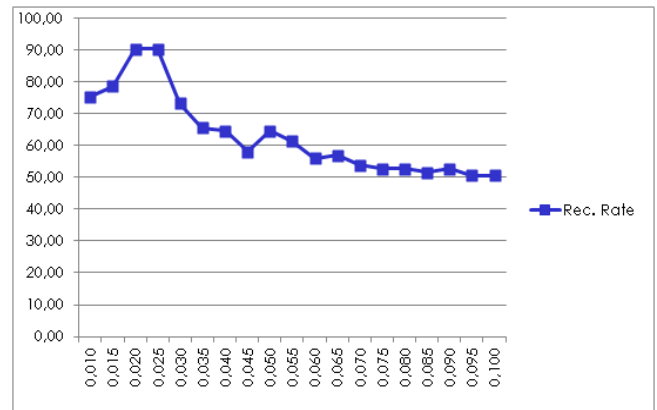


Figure 5. Threshold estimation in segment representation for Weizmann data set

The comparison of our method in terms of recognition rates with the essential studies having different viewpoints in human action recognition and segment classification is shown in Table 3.

Table 3. Comparison of the results of video segment classification for Weizmann data set

METHODS	RECOGNITION RATES
Chaudry et al. 2009 [35]	94.44%
Ali et al. 2007 [33]	92.60%
This Paper	90.32%
Niebles et al. 2008 [30]	90.00%
Lertniphonphan 2011 [15]	79.17 %
Niebles et al. 2007 [8]	72.80 %

The methods shown in the table are some of the well-known reference studies tackling with the temporal segment classification problem. The approach in [33] uses new interest point features having time dimension. The classification is done according to these novel 3D features. [35, 15] proposes optical flow based classification, whereas [35] focuses on representing the segments frame by frame optical flows with high dimensions causing curse of dimensionality problem. Instead of dealing with the representation, the method aims to contribute by finding new metrics and time series patterns in this high dimensional data. [15], on the other hand, proposes a representation structure based on direction histograms of optical flow. [30] presents a model based on spatio-temporal words. This method sees the segments as bag-of-features

and makes the classification according to the code words. A bag-of-features method is also proposed in [8] where interest points are extracted and used as bag-of-features.

When we analyze the results, we have seen that the methods proposing interest point based new 3D features are more successful than the other models. But, the features are specific to the data set which makes the solution dependent on data set types. Methods focusing on mining the highly over-descriptive data in terms of time domain exhibit high success rates as they develop the model independent from the contributions in video features. But, they are disadvantageous with their high dimensional representation regarding time complexity. Our optical flow based method better results than the approaches using optical flow based segment representation. It is also more successful than the bag-of-words based methods.

7 Conclusion

In this study, we tried to solve a combination of different problems on action recognition. The fundamental problem inspires us is the representation of temporal information. In many fields, representation of temporal information is essential to retrieve information from a temporal data set. The solution to the problem varies from representing each temporal entity in a different time slice to representing a simple summary of the whole time interval. Efforts for finding a solution between these two endpoints, should try to tackle the problem from different point of views. This is because, the level of representation changes with the source of the problem. For instance, to represent all the information in all time slices for symbolizing the temporal information having high frequency over time, one should handle the curse of dimensionality problem. On the other hand, representing a single summary will cause the problem of lacking the flow of temporal information. In these cases, the focus of approaches will be finding supportive information from different sources and integration of these sources in a singular representation.

We aimed to solve the temporal information representation problem in video domain. As the video information is a perfect example of high frequency temporal information, representation of video information is essential for the purposes based on video information retrieval. Video action recognition is selected as our specific domain. The problem domain is reduced to the temporal video segment classification. The study is shaped on visual features of the video information for the automaticity concerns. As it is mentioned below, the representation level determines the reduced problem. In this context, our aim is to represent the video scenes avoiding the lack of temporal information flow while without causing the curse of dimensionality problem. Therefore, using more descriptive and high level visual features having the ability to host the additional temporal nature of the simpler features such as color, edge, corner, etc. becomes unavoidable. This will pass the high load of temporal information residing in high dimensional representation to the mentioned high level features.

The discussion summarized above took us to the complex visual features having temporal dimension. In our research, we observed that space-time related 3D features obtained from combining 2D features with temporal

information [34, 39] are proved to be successful. Space-time interest points and space-time shapes for actions are proposed in these studies. We also observed high level state-of-the-art features such as optical flow describing the motion of frame features are calculated and used in temporal video information representation as in [15, 35, 38]. Curse of dimensionality problem occurs as all frames are represented using optical flow vectors [35]. The problem is solved by using time series analysis and metrics.

An optical flow based approach is proposed in this paper for representing temporal video information by inspiring from the above studies. This generic approach is applied in both temporal video segment classification and temporal video segmentation. The adaptation of the model to video segment classification is presented. The weighted frame velocity concept is proposed to strengthen the representation with the velocity of video frames. This representation formalism is tested with SVM based classification of video segments. The results show that the proposed method produces encouraging results.

The main advantage of the method is the multi-purpose temporal video representation model proposed for video action recognition domain. The new formalism described here is especially important for simplifying the computational complexity for high dimensional information.

8 References

- [1] T. C. Vasileios, C. L. Aristidis and P. G. Nikolaos, "Scene Detection in Videos Using Shot Clustering and Sequence Alignment". *IEEE Transactions on Multimedia*, vol. 11, no. 1, January 2009.
- [2] A. Ghoshal, P. Ircing and S. Khudanpur, "Hidden Markov Models for Automatic Annotation and Content Based Retrieval of Images and Video". *Proc. of SIGIR*, 2005.
- [3] L. W. Chang, W. N. Lie, and R. Chiang, "Automatic Annotation and Retrieval for Videos". *Proc. of PSIVT 2006*, LNCS 4319, pp. 1030 – 1040, 2006.
- [4] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video Event Classification Using String Kernels". *Multimed Tools Application*, 48:69–87, 2009.
- [5] F. Wang, Y. Jiang and C. Ngo, "Video Event Detection Using Motion Relativity and Visual Relatedness". *ACM Multimedia '08*, October 26–31, 2008,
- [6] C. Ngo, T. Pong and H.Zhang, "Motion-Based Video Representation for Scene Change Detection". *International Journal of Computer Vision* 50(2), 127–142, 2002
- [7] P. Sand and S. Teller, "Particle Video: Long-Range Motion Estimation Using Point Trajectories". *International Journal of Computer Vision*, 2008.
- [8] J. C. Niebles and L. Fei-Fei, "A Hierarchical Model of Shape and Appearance for Human Action Classification". In *IEEE International Conference on Computer Vision and Pattern Recognition*, June 2007.
- [9] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies". *Proc. of CVPR'08*, Anchorage, US, 2008.

- [10] A. Burton and J. Radford, "Thinking in Perspective: Critical Essays in the Study of Thought Processes". Routledge, 1978.
- [11] D. H. Warren and E. R. Strelow, "Electronic Spatial Sensing for the Blind: Contributions from Perception". Nato Science Series.
- [12] J. J. Ibson, "The Perception of the Visual World". Houghton Mifflin, 1950.
- [13] C. S. Royden and K. D. Moore, "Use of speed cues in the detection of moving objects by moving observers". *Vision research*, 59, 17–24, 2012.
- [14] C. Gianluigi and S. Raimondo, "An innovative algorithm for key frame extraction in video summarization". *Journal of Real-Time Image Processing*, vol. 1, no. 1, pp. 69–88, 2006.
- [15] K. Lertniphonphan, S. Aramvith and T. Chalidabhongse, "Human Action Recognition using Direction Histograms of Optical Flow". In *ISCIT*, 2011.
- [16] J. J. Gibson, "The Perception of the Visual World". Houghton Mifflin, 1950.
- [17] J. Barron, D. Fleet and S. Beauchemin, "Performance of optical flow techniques". *International Journal of Computer Vision*, pp. 43-47, 1994.
- [18] T. Camus, "Real-time quantized optical flow". *The Journal of Real-Time Imaging (special issue on Real-Time Motion Analysis)*, 3:71–86, 1997.
- [19] B. K. P. Horn and B. G. Schunck, "Determining optical flow". *Artificial Intelligence*, no. 17, pp. 185-203, 1981.
- [20] M. Proesmans, L. Van Gool, E. Pauwels and A. Oosterlinck, "Determination of optical flow and its discontinuities using non-linear diffusion". In *3rd European Conference on Computer Vision, ECCV'94*, vol. 2, pp. 295–304, 1994.
- [21] P. Anandan, "A Computational Framework and an algorithm for the measurement of visual motion". *International Journal of Computer Vision*, no. 2, pp. 283-310, 1989.
- [22] B. Buxton and H. Buxton, "Computation of optical flow from the motion of edge features in image sequences". *Image and Vision Computing*, no. 2, pp. 59-74, 1984.
- [23] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information". *International Journal of Computer Vision*, no. 5, pp. 77-104, 1990.
- [24] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision". *Proc. of DARPA IU Workshop*, pp. 121-130, 1981.
- [25] H. H. Nagel, "On the estimation of optical flow relations between different approaches and some new results". *Artificial Intelligence*, no. 33, pp. 299-324, 1987.
- [26] D. J. Heeger, "Optical flow using spatiotemporal filters". *International Journal of Computer Vision*, no. 1, pp. 279-302, 1988.
- [27] A. Singh, "An estimation-theoretic framework for image-flow computation". *Proc. of ICCV*, pp. 168-177, Osaka, 1990.
- [28] A. M. Waxman, J. Wuand and F. Bergholm, "Convected activation proles and receptive fields for real time measurement of short range visual motion". *Proc. of IEEE CVPR*, pp. 717-723, Ann Arbor, 1988.
- [29] S. Uras, F. Girosi, A. Verri and V. Torre, "A computational approach to motion perception". *Biol. Cybern.*, no. 60, pp. 79-97, 1988.
- [30] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words". *International Journal of Computer Vision*, 79:299–318, 2008.
- [31] C. Harris and M. Stephens, "A combined corner and edge detector". *Proc. of the 4th Alvey Vision Conference*. pp. 147–151, 1988.
- [32] J. Shi and C. Tomasi, "Good features to track". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, June 1994.
- [33] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition". In *IEEE International Conference on Computer Vision*, 2007.
- [34] I. Laptev and T. Lindeberg, "Space-Time interest points". *Proc. of ICCV'03*, Nice, France, pp.I:432-439, 2003.
- [35] R. Chaudry, A. Ravichandran, G. Hager and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions". *Proc. of CVPR'09*, Miami, US, 2009.
- [36] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance". In *IEEE International Conference on Computer Vision*, pp. 726–733, 2003.
- [37] D. Tran and A. Sorokin, "Human activity recognition with metric learning". In *European Conference on Computer Vision*, 2008.
- [38] F. Erciş, "Comparison of histograms of oriented optical flow based action recognition methods". MS Thesis, Middle East Technical University, Turkey, 2012.
- [39] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.
- [40] S. Little, I. Jargalsaikhan, K. Clawson, M. Nieto, H. Li, C. Direkoglu, N. E. O'Connor, A. F. Smeaton, B. Scotney, H. Wang and J. Liu, "An information retrieval approach to identifying infrequent events in surveillance video". *Proc. of the 3rd ACM Conference on International Conference on Multimedia Retrieval (ICMR '13)*, ACM, New York, NY, USA, 2013.
- [41] T. Wang and H. Snoussi, "Histograms of optical flow orientation for abnormal events detection". *Performance Evaluation of Tracking and Surveillance (PETS)*, 2013 *IEEE International Workshop on*, vol., no., pp.45,52, 2013.

Gesture recognition in cooking video based on image features and motion features using Bayesian Network classifier

Nguyen Tuan Hung¹, Nguyen Thanh Binh², Pham The Bao¹, and Jin Young Kim³

¹Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Viet Nam

²Faculty of Management Information System, College of Finance and Customs, Ho Chi Minh City, Viet Nam

³Chonnam National University, Gwangju, Korea

Abstract - *This paper proposes a method combining image features and motion features for gesture recognition in cooking video. By using image features including global and local features of RGB images and then representing those using bag of features, motions in video are represented. We also use relative positions between objects after they are detected in this frame. Motions are also represented through motion features calculated from frame sequences using dense trajectories. After all, we combine both image features and motion features to describe cooking gestures. We use Bayesian Network model to predict which action class a certain frame belongs to base on the action class of previous frames and the cooking gesture in current frame. Our method has been approved through ACE dataset that it can recognize human cooking action as we expected. In addition, it is also a general method for solving cooking actions recognition problem..*

Keywords: action recognition; Bayesian network; features combination; image features; motion features; depth image

1 Introduction

Cooking and eating are daily activities which all of us have to do in order to stay healthy. Although these are simple tasks that anyone would have to go through every day in life, they account for a very important position because of healthy impact. On the other hand, in a modern society, time has become more precious than ever before. Everyone does not have much time for cooking. It leads to a direct impact on the health of everyone. Therefore, a question “How could we have a delicious and nutritious dish with less cooking time?” has been raised for a while.

In recent years, researchers all over the world have been building various intelligent kitchen systems, which are the answers for the above question. These systems can help users cook faster and more efficiently. In these systems, solutions of many different problems such as object recognition problem, human action recognition, or nutritious meals computation are combined. All of the above problems have been actually raised in the “Multimedia for Cooking and Eating Activities” workshop from 2009. Among these problems, we evaluated

that the human’s cooking action recognition is the most challenge problem. Many complex challenges still exist and there is not any complete solution until now.

One of its challenges is action recognition problem. Its object is how a computer program can recognize cooking actions based on training data. Furthermore, based on sequences of cooking actions, it could predict what kind of dishes. In practice, when this program is being executed, it observes actions of user(s), recognizes these actions, and either warns user(s) if there is any wrong or suggests next cooking steps. Therefore, we realize that solving problem of cooking action recognition is the most important task to complete our intelligent kitchen system. This problem has been mentioned in a contest [1] of ICPR2012 conference to find out the solutions from many researchers. A new “Actions for Cooking Eggs” (ACE) Dataset [2], which we used for evaluating our method in experiments, was presented in this contest.

In this paper, we propose a method combining image features and motion features for gesture recognition in cooking video. We divide this problem into four sub-problems that are cooking action representation by image features, cooking action representation by motion features, combination of image features and motion features, and cooking action classification. From a cooking video, firstly, we need to represent the cooking actions by extracting some image features such as PHOG [3], or SIFT [4] and motion features such as dense motion [5]. We also detect objects on each frame and compute the relative positions between them. If we can represent actions more exactly, classification result will be better. Hence, this step is very important and extracted features must be chosen carefully. Another sub-problem is to combine these features because using only one kind of feature is not good enough in accuracy. In case combining different features, accuracy gets higher. However, the way of combination is still a complex problem since it depends on datasets and kinds of feature. In this paper we use both early fusion and late fusion techniques. [6, 7]

Last but not least, to solve cooking action classification problem, we use Bayesian network classifier. The first reason for choosing Bayesian network classifier is that our method

based on the features of both current frame and previous frames. In cooking video, the sequence of actions for each kind of dishes has its characteristics. So we can learn this feature and use it in classifier to achieve better classification results. This is the second reason for us to choose Bayesian network classifier. Another advantage of this classifier is easy to update parameters of nodes in the network and also easy to modify by adding or removing nodes in the network, which is the third reason for using Bayesian network. For three main reasons, we have chosen Bayesian network as the classifier for our system.

To build a gesture recognition system, our main contributions are as follows:

- Representing cooking motion by image features including color histogram, LBP, EOH, PHOG, SIFT and also the relative positions of objects.
- Representing cooking motion by motion features using dense trajectories motion feature.
- Combination of image features and motion features by both early fusion and late fusion techniques.
- Using Bayesian network classification for gesture recognition.

2 Related Work

Human's action recognition problem has appeared for a long time in many applications especially for smart devices. In general, the human's action recognition systems work by extracting some kind of features and combining them in a certain way. These systems usually use both global and local image features. Many researchers have been trying to answer which feature is the best for describing human's action and whether different features are supplements for each other. Throughout the recent years, these following features have been studied to answer the above questions.

One answer came from a successful system built by a research team from Columbia University. It used SIFT [4] as an image feature, STIP [8] as a motion feature and MFCC (Mel Frequency Cepstral Coefficients) [9] as a sound feature. Overall, STIP was the best motion feature for human's action description. However, to achieve better results, different kinds of them, which were supplemental features, should be combined together including image features, motion features, and even sound features. This is an important conclusion that other teams agree with.

Researchers from IBM built another human's action recognition system [10]. It used many image features including SIFT [4], GIST [11], Color histogram, Color Moment, Wavelet Texture, etc. For motion features, it uses STIP [8] combining with HOG/HOF [12]. According to their experiment's result, they concluded that a combination of some features raised the accuracy of recognition. This

conclusion is the same as Columbia team [13]. Besides, Nikon team's system is a simple system [14]. It used scene cut detection to extract key frames. This method depends on kind of videos, and length of videos. Although this method is not the best method, it is a good idea that we can use in our system.

In ICPR 2012 conference, there were 6 systems submitted to. To describe their cooking action system, researchers from Kyushu University, Japan [15] used local features including FAST detector and CHOG3D descriptor, and combined with hand motion feature which was extracted from depth images. In that research, they achieved some spectacular results. The average accuracy for action recognition in their experiment is 50.6% in case of using depth local feature and 53.0% when they used local feature. Then, when they combine these kinds of feature, the accuracy is achieved 57.1%.

The winning entry was by a team from Chukyo University, Japan. Their method uses heuristic approach using image features with some modifications. Then, in post-processing step, they use some methods to avoid unnatural labeling results. Their result proved that this approach is more effective than other approaches in practical use. Besides, the other recognition systems from other teams are also interesting. They use Motion History Image feature, Spatio-Temporal Interest Point Description feature, trajectories feature, and context information. Moreover, they use one-vs-all linear SVM classifier as an action model. In post-processing, they apply 1D Markov Random Field on the predicted class labels.

According to these methods, we could conclude that an effective action recognition method usually uses some different features and combines them to achieve the better result. Besides, the information such as context information and cooking action sequence are also important for improving the accuracy of our method.

3 Our Method

3.1 Our recognition system overview

There are two diagrams, which are training framework and testing framework of our recognition system. As depicted in Fig. 1, main steps in these frameworks are the same. First of all, the input videos are pre-processed including calibrations, table area and floor area segmentation, objects and human detection, and hands detection. After that, some features are extracted from both depth images and color images. According to original action recognition methods, they extract motion features in the videos and perform discriminative analysis for action recognition. In cooking action recognition problem, however, some actions are not described by using motion feature alone. So, we propose to extract both human's motion feature and image features from cooking video.

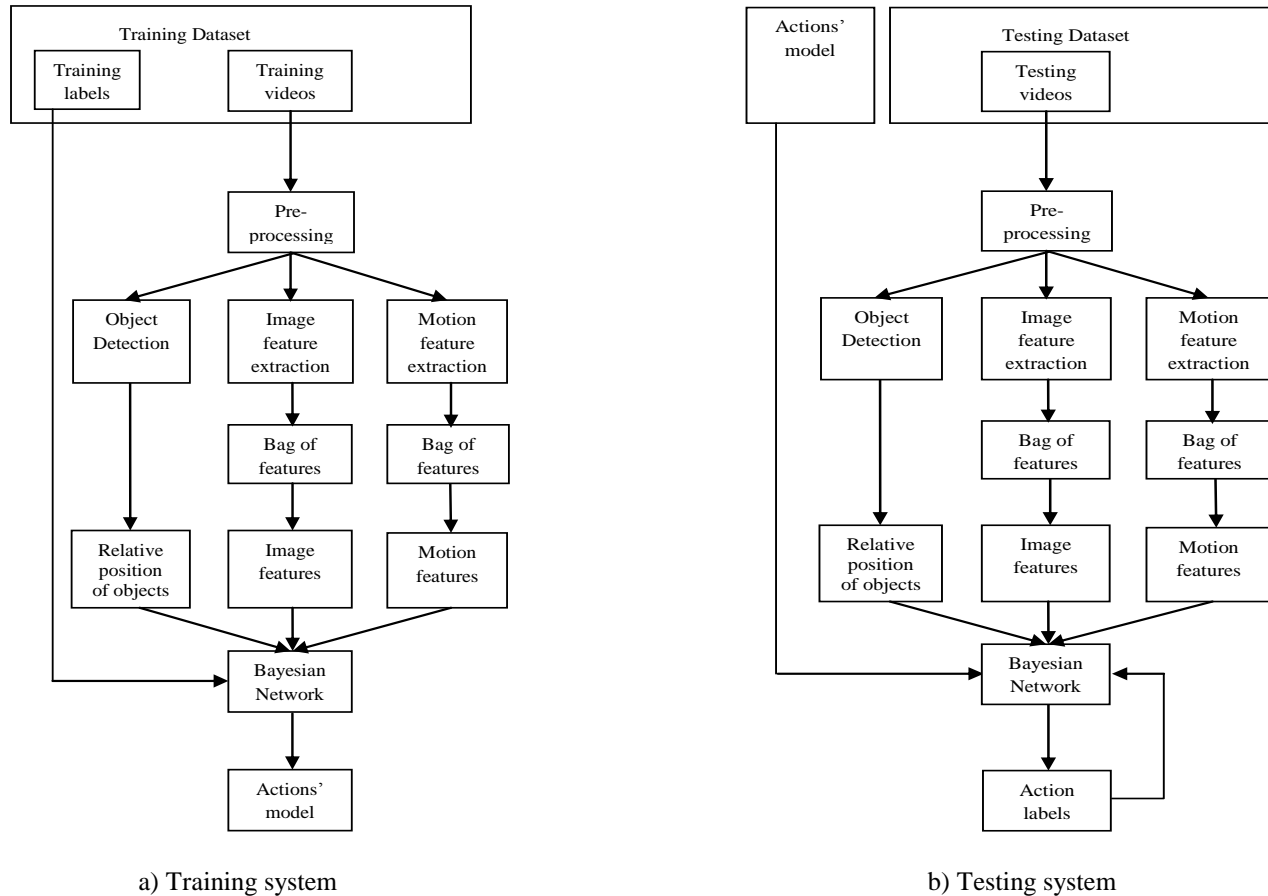


Fig. 1. Diagrams to describe our method's systems (a) is our training system and (b) is our testing our testing system.

To match the feature vectors, we can use some matching method. However, in this problem, we have a large of feature vectors, which we cannot use original matching methods. Hence, we use Bag of Feature (BoF) [16] to describe image features and motion features for speeding up the matching step. We also detect objects from frames and compute their relations. Moreover, from the training labels, we can learn some rules about the sequences of actions. Next steps are Bayesian Network (BN) construction and parameters learning.

We create three Bayesian networks, and train their parameters. The output of training system is a model which describes the cooking actions. In the testing system, we use the trained model to classify cooking actions. Besides, we also use action labels of the previous actions as an input data for BN classifier. The action label of each frame is the output of testing system. Then, we evaluate our method based on this output. For evaluating, we calculate accuracy score from precision and recall manner¹. Their harmonic, after that, is calculated by following formula

$$F = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \right)} \quad (1)$$

The final score is given by averaging all F-measures of all cooking motion label.

3.2 Processing input data

In preprocessing step, we have prepared data for available to use in the following steps.

First of all, we need calibrate depth images because of a distance separates depth camera and color camera. The Kinect disparity is related to a normalized disparity

$$d = \frac{1}{8} * (d_{\text{off}} - k_d) \quad (2)$$

Where d is a normalized disparity, k_d is the Kinect disparity, and d_{off} is an offset value particular to a given Kinect device. According to a technical Kinect calibration report [17], d_{off} is typically around 1090. In depth image, there is a small black

¹ <http://www.murase.m.is.nagoya-u.ac.jp/KSCGR/evaluation.html>

band on the right of the image. We expect to eliminate them for mapping the depth image to color image later.

We use depth images to segment table area and floor area. Some first frames are chosen to find a border that separates table and floor areas because there is no human in these frames. Therefore, it is easy to find this border. Sum of grayscale of the depth image for each row is calculated and then ratio of disparity of two adjacent rows is calculated, too. In the case that this ratio is higher than a threshold, this row is the border separated table and floor areas.

For human detection, we use depth image and floor area, which is located in previous step. As we known, when human appears in scene, the distance from depth sensor to human body is much nearer than floor area. Therefore, the pixels, whose grayscale are higher than average grayscale of floor area, are ones of human body.

Next, for hand detection, we use color images and apply skin detection. First step, we choose some points, which are in skin area. Based on the color of these points, we obtain the range of the color of pixels in skin area. Then, each pixel is classified by its color. Then, we eliminate areas that are not in range [hand_size_min, hand_size_max], which are the thresholds to determine size of hands. If there are still more than two areas, only two largest areas are selected.

Objects in this case mean cooking tools such as fry pan, pan, chopstick; and ingredients such as egg, ham, and seasoning. For object detection, we use color images because they contain more information than depth images. First, we collect the image samples of each object. Each kind of object has about 100 images. Next, image feature for each of image is calculated and a training data is made using one-vs-all SVM classification. When testing system is executed, a slide window is used to detect objects and recognize which kind of object.

3.3 Image Feature Extraction

After preprocessing step, firstly, image features are extracted from frames of cooking video. However, we only extract feature from key-frames which are chosen by one frame for each k frames in video. In our experiment, we use k is 10 to reduce the amount of computation because in ten continuous frames there are not much differences. In our research, Local Binary Pattern (LBP) [18], Edge Oriented Histogram (EOH) [19], Pyramid Histogram of Oriented Gradient (PHOG) [3], and Scale Invariant Feature Transform (SIFT) [4] are used because they can characterize the content of frames including information about the context and are easy to be extracted.

In addition, each image feature is extracted from cells of a 4×8 grid of frame. For PHOG feature, we extract with 8 gradient bins and the highest level $l = 2$. Moreover, for SIFT features, we apply BoF method [16] to increase the effectiveness of

recognition. After the step of keypoint detection, histogram of these keypoints is calculated based on codebook, which is the collection of millions of keypoints. We gain many different kinds of feature vectors. Then, we apply early fusion technique in [6] to join them together to obtain only one feature vector, which is called the image feature vector characterized for a frame in video.

3.4 Motion Feature Extraction

Being parallel with the image features extraction, we extract motion features from videos. In our method, we use dense trajectories and motion boundary histogram (MBH) description [5] for action representation. The main reason for choosing this motion feature is every cooking actions are characterized by different simple motions, such as cutting action is related to vertical motions while mixing action is almost described by turn around motions. Moreover, there are many fine motions in cooking videos, so that we use dense trajectories feature that is the best feature for representing even the fine motions. Besides, MBH descriptor expresses only boundary of foreground motion and eliminates background and camera motion. Thus, it is completely appropriate to be applied in this step for action representation.

To compute the optical flow from above dense samples, we use Farneback algorithm [20] because it is one of the fastest algorithm to compute a dense optical flow. Next, we track in optical flows to find out trajectories in a sequence of 15 continuous frames. To describe motion feature, each video is separated to many blocks, which are $N \times M \times L$ -size blocks. It means scaling each optical flow matrix to size $N \times M$ and each block containing L optical flow. Then, each block is divided into $n_\sigma * n_\sigma * n_t$ cells. Lastly, we calculate MBH feature for each cell and join them together. For motion feature, we also use Bag of features [16] to increase effectiveness of recognition as SIFT feature from image features.

3.5 Bayesian Networks Training

According to three main reasons in the first section, we choose Bayesian networks as our classifier. In our approach, we use three separate networks that play different role in this classification step. Since there are some categories of features from label information, image features and motion features, we use different Bayesian networks for training and classifying each of categories. By using three different Bayesian networks, the classification result would be better than using only one network. Moreover, we can train three different networks at the same time which means training time could be reduced.

In this step, we create three BNs to classify feature vectors into a certain action class. First of all, we have a BN from training label data which represents the possibility of subsequence action's label based on previous identified action labels. It is calculated by using Bayes's theorem formula

$$P_{\text{NEXT}}(A_i) = P(A_i | \text{PALs}) \quad (3)$$

where A_i is the i^{th} action and PALs are previous action labels.

For the second BN, we have a graph in which nodes' value are extracted from high-level feature. In this BN, see fig. 2, we have Human node which determines whether human exists in this frame. Similarly, Hand node is node which determines whether hands are in frame or not. Besides, nodes including Tool Using node, Container Using node, and Food Using node are determined based on the relative position between the hands and the objects. In addition, there is a Status Changing node, which expresses the changing of ingredient inside cooking container. Finally, the action label nodes are based on the above-identified nodes, whose conditional probability formula is below

$$P_{\text{IMAGE}}(A_i) = P(A_i | SC, TU, CU, FU) * P(CU | Hd) * P(FU | Hd) * P(TU | Hd) * P(SC | Hm) * P(Hd | Hm) * P(Hm) \quad (4)$$

where SC is Status Changing, TU is Tool Using, CU is Container Using, FU is Food Using, Hd is Hand and Hm is Human.

The last BN represents the possibility of an action based on motion features of sequence of images surround the current frame.

$$P_{\text{MOTION}}(A_i) = P(A_i | MF) \quad (5)$$

where MF is motion feature.

The probability of the i^{th} action class calculated by the first BN is called $P_{\text{NEXT}}(A_i)$, by the second BN is called $P_{\text{IMAGE}}(A_i)$, and by the last one is called $P_{\text{MOTION}}(A_i)$. We can compute the probability of an action label based on P_{NEXT} and P_{IMAGE} only by multiply probabilities, or P_{NEXT} and P_{MOTION} . After that, we combine these probability values to obtain the probability of the i^{th} action

$$P(A_i) = w_1 * P_{\text{NEXT}} * P_{\text{IMAGE}} + w_2 * P_{\text{NEXT}} * P_{\text{MOTION}} \quad (6)$$

4 Experiments

4.1 Dataset

In our experiments, we use ACE dataset, which contains five sets for training and two sets for testing. There are five menus of cooking eggs and eight kinds of cooking actions performed by actors in dataset. In addition, the ingredients and cooking utensils, which are used in dataset, are egg, ham, milk, oil, salt and frying pan, saucepan, bowl, knife, chopsticks... The videos were captured by a Kinect sensor, which provides synchronized color and depth image sequences. Each of the videos was from 5 to 10 minute long containing from 2,000 to over 10,000 frames. Each frame is 480*640-size and is assigned to a certain action label indicating type of action performed by the actors in video.

In this dataset, all dishes are based on egg, sometimes ham or some seasons are added to. Each dish has its own color such as boiled egg has white or brown color from eggshell color while omelet has yellow and pink color from egg and ham. Therefore, we used image features such as color histogram,

color moment feature which are extracted to classify different dish. Besides, because each cooking action requires different cooking tool, which has characterized shape, we use image features related to edge features such as cutting action requires knife while mixing action requires chopsticks.

4.2 Parameter setting

There are some parameters throughout our processes such as in preprocessing step, the parameter $d_{\text{off}} = 1090$ which depends on a certain Kinect device. Other parameters are the thresholds determining hands are `hand_size_min` and `hand_size_max` which are obtained from training data. For motion extraction, there are also some other parameter including $N * M = 480 * 640$, $L = 20$, $n_c = 2$ and $n_t = 3$. Lastly, we simply use $w_1 = w_2 = 1$ in (6) because the problem of optimizing value for w_1 and w_2 is hard problem. However, by using $w_1 = w_2 = 1$ we also obtain a good result as we expected.

4.3 Results

We evaluate the recognition precision of image features, motion features and combination of them in ACE dataset. The evaluation results of using either image features or motion features singly are shown in the first and the second columns of Table 1. When using only image features, some actions like boiling, breaking, seasoning, etc. cannot be classified, which precision is 0%. While other actions include baking action and cutting action achieved has better precisions which are 36.9% and 26.1%. However, when we apply only motion features, all actions have better precision than using only image features.

Therefore, we find out that the motion features are more efficient to represent the cooking actions than the image features, especially for the actions with large amplitude such as baking or cutting using a big cooking tool as a knife. It is successful to recognize based on the image features. However, they do not describe the other actions which have small

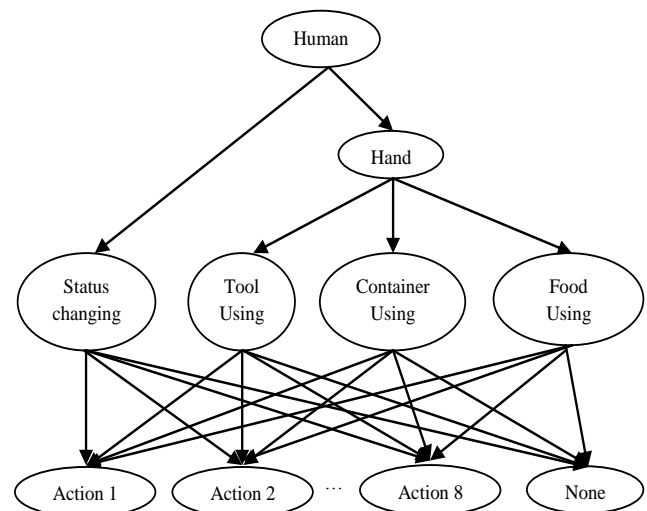


Fig. 2. The second Bayesian Network in our method.

amplitude or use tiny cooking tools. On the other hand, when we use only image features, the actions with no color changing or no use any cooking tools such as boiling, breaking or seasoning could not be recognized by image features including color feature or edge feature. After that, we combine both of image features and motion features to obtain better results as we have expected.

The result of a combination is shown in the last column of Table 1. Compared with two first cases, the combination of the two features improves the recognition results. In this case, some actions are also well recognized, especially the actions “baking” and “turning”, because they are the ones with large amplitude. However, some action recognitions are not improved while comparing with motion features because image features may not describe exactly these actions. To solve the problems, one way is to modify the high-level image feature extraction to describe more exactly such as determining exactly which cooking tools are being used. The other way is to combine these features in more efficient way such as trying to find better parameters w_1 and w_2 .

Moreover, the recognition results of each testing video are shown in table 2. Each video contains a sequence of actions to cook a certain dish. In our experiment, the best precision is over 40% for all frames in a video. Although the average of precision is approximate 30%, it proves that our approach is appropriate. If we optimize the parameters and improve our implementation, the result would be definitely improved. Therefore, our works still need to be continued in near future.

In our research, we use a PC with CPU Intel core i7-2600 3.4GHz, 8 GB RAM, and MATLAB 7.12.0 (R2011a) 64-bit on Windows 7 64-bit OS. Most of time is used to extract motion features although we have applied one of the fastest optical flow computing algorithms. It requires much more running time because dense trajectories in each sequence of images are extracted. In addition, we only use one PC in our research. If our method is applied on a parallel system such as the clusters or the high performance computing using GPU-CPU, the result will be improved more.

5 Conclusions

In this paper, we proposed a method using both image features and motion features for gesture recognition in cooking video. It means the motions in cooking video are represented by image feature vector and motion feature vectors. In our method, Bayesian Network model is model to predict which action class a certain frame belongs to based on the action class of previous frames and cooking gesture in present frame. Additional information such as the sequence of actions is also applied into Bayesian network model to improve classification result.

According to our results, our proposed method is a good approach for solving action recognition in video. Although its performance is not good enough when comparing with the best method, we are certain this method can be improved to

achieve higher performance. In addition, it is a completely flexible method as we can add easily more action or other features. Furthermore, we can also reconstruct Bayesian networks and update their parameters in nodes easily too. Thus, our method can be applied for other action recognition systems even there are many complex actions.

In the future, we are going to improve motion feature extraction, which is acceleration of feature extraction because now it account to over 80% of the running time. Another problem that we can improve in near future is using high level features. In our research, at present, there is still limitation in high level features application because they still require more computation and time now.

TABLE I. ACTION RECOGNITION PRECISION (%)

Action	Image features	Motion features	Combination
breaking	0.0	12.7	15.0
mixing	0.0	22.0	22.5
baking	36.9	38.4	41.9
turning	0.0	54.5	54.6
cutting	26.1	17.3	17.4
boiling	0.0	29.2	29.1
seasoning	0.0	15.9	16.0
peeling	0.0	27.5	27.5
Average	7.9	27.2	28.0

TABLE II. RECOGNITION PRECISION (%) ON TESTING VIDEOS

Video No.	Image features	Motion features	Combination
1	24.4	25.0	26.3
2	8.3	41.6	43.3
3	12.2	23.1	24.2
4	7.1	16.1	15.9
5	6.0	29.1	30.1
6	20.5	21.2	23.5
7	7.2	15.7	17.1
8	0.0	18.4	18.5
9	4.8	27.1	28.2
10	12.3	32.8	33.9

Acknowledgement

We would like to thank Atsushi Shimada, Kazuaki Kondo, Daisuke Deguchi, Géraldine Morin, Helman Stern for KSCGR contest organization and Tomo Asakusa et al. from Kyoto University, Japan for creating and distributing Actions for Cooking Egg dataset.

6 References

- [1] ICPR 2012 Contest, Kitchen Scene Context based Gesture Recognition <http://www.murase.m.is.nagoya-u.ac.jp/KSCGR/index.html>
- [2] Atsushi Shimada, "Kitchen Scene Context Based Gesture Recognition: A Contest in ICPR2012," *Advances in depth image Analysis and Applications*, 2013
- [3] Bosch A., Zisserman A., and Munoz X., Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.
- [4] Lowe D. G., "Distinctive image features from scale-invariant keypoints," in *Int. J. Comput. Vision*, Nov. 2004, vol. 60, pp. 91-110.
- [5] Wang H., Klaser A., Schmid C., and Liu C. L., "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, June 2011, pp. 3169-3176.
- [6] Snoek C. et al., The MediaMill TRECVID 2004 semantic video search engine. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2004.
- [7] Westerveld T. et al., A probabilistic multimedia retrieval model and its evaluation. *EURASIP JASP*, 2003.
- [8] Ivan Laptev, "On space-time interest points," in *International Journal of Computer Vision*, Sept. 2005, vol. 64, pp. 107-123.
- [9] P. Mermelstein (1976), "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligent*, C. H. Chen, Ed., pp. 374-388. Academic, New York.
- [10] Matthew H., Gang H., Apostol N., John R. S., Lexing X., Bert H., Michele Merler, Hua Ouyang, and Mingyuan Zhou, "IBM research trecvid-2010 video copy detection and multimedia event detection system," in *NIST TRECVID Workshop*, 2010.
- [11] Aude Oliva and Antonio Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," in *International Journal of Computer Vision* 2001, vol.42, pp. 145-175.
- [12] Dalal N. and Triggs B., "Histograms of oriented gradients for human detection," in *CVPR (1)*, 2005, pp. 886-893.
- [13] Yu G. J., Xiaohong Z., Guangnan Y., Subhabrata B., Dan E., Mubarak S., and Shih F. C., "Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching," in *NIST TRECVID Workshop*, 2010.
- [14] Takeshi Matsuo and Shinichi Nakajima, "Nikon Multimedia Event Detection System" in *NIST TRECVID Workshop*, 2010.
- [15] Ji Y., Ko Y., Shimada A., Nagahara H., and Taniguchi R., "Cooking Gesture Recognition using Local Feature and Depth Image. ICPR 2012.
- [16] Cruska G., Dance C. R., Fan L., Willamowski J., and Bray C., "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 886-893, 2004.
- [17] Bueno, J.G.; Slupska, P.J.; Burrus, N.; Moreno, L., "Textureless object recognition and arm planning for a mobile manipulator," *ELMAR, 2011 Proceedings*, vol., no., pp.59,62, 14-16 Sept. 2011.
- [18] Ojala T., Pietikainen M., Harwood D., "A comparative study of texture measures with classification based on feature distributions", *Pattern Recognition* vol. 29, 1996.
- [19] Freeman W. T., Michal R., "Orientation Histograms for Hand Gesture Recognition", *IEEE Intl. Wkshp. on Automatic Face and Gesture Recognition*, Zurich, June, 1995.
- [20] Farneback G., "Two-frame motion estimation based on polynomial expansion", *Lecture Notes in Computer Science*, 2003, (2749), 363-370.

Improving Performance using both of Correlation and Absolute Difference on Similar Play Estimation

Kyota Aoki and Ryo Aita

Graduate school of Engineering, Utsunomiya University, Utsunomiya, Tochigi, JAPAN

Abstract - Plays in sports are described as the motions of players. This paper proposes the similar play retrieving method based on the motion compensation vector in MPEG sports videos. In MPEG videos, there are motion compensation vectors. Using the motion compensation vectors, we don't need to estimate the motion vectors between adjacent frames. This work uses the 1D degenerated descriptions of each motion image between 2 adjacent frames. Connecting the 1D degenerated descriptions on time direction, we have the space-time image. This space-time image describes a sequence of frames as a 2-dimensional image. Using this space-time image, this work shows the performance using both of the correlation and the absolute difference to retrieve a small number of plays in a huge number of frames based on a single template play. Our experiment records 0.94 as recall, 0.800 as precision and 0.865 as F-measure in 138 plays in 132503 frames.

Keywords: play estimation; motion compensation vector; MPEG video; Absolute difference; correlation

1 Introduction

There are many videos about sports. There is a large need for content-based video retrievals. The amount of videos is huge, so we need an automatic indexing method [1][2][3][4][5]. We proposed the method retrieves shots, including a similar motion based on the similarity of the motion with a sample part of videos [6][7][8][9].

In former works, we used the correlation of motions and the correlation of textures [10]. We have a good performance using both of the correlation of motions and the correlation of textures.

This paper proposes the method to retrieve the plays using only motion compensation vectors in MPEG videos. Using multiple features, the performance increases. Many works try to index sport videos using the motions in the videos. Many works try to understand the progress of games with tracking the players. Many of the works use the motion vectors in MPEG videos. They succeed to find camera works. They are zoom-in, zoom-out, pan and etc. However, no work retrieves a play of a single player from only motions directly. Of course, camera works have an important role in understanding videos. Sound also has some roles in understanding videos. Many works use camera works and sound for understanding sport videos. Those feature-combining methods have some

successes about retrieving home-runs and other plays. However, those works did not success to retrieve plays from only motions. Recently, many works focused on retrieval combining many features and their relations [11][12].

In this paper, we try to propose the method that retrieves similar plays only from motions. With the help of texture, the performance of similar play retrieval increases. We showed that the combination both of motions and textures shows better performance about similar play retrieval [13]. However, in many games, there are changes of fields and spectators. The colors of playing fields changes with the change of seasons. The motion based method can work with textures, sound, and camera works. However, this paper proposes the method to estimate similar plays only from motions in motion compensation vectors in MPEG videos. There are many motion estimation methods [14][15]. However, they need large computations. The method gets the motions from motion compensation vectors in MPEG videos, and makes the 1-dimensional projections from the X motion Y motion. The motion compensation vector exists at each 16x16 pixels square. It is very sparse description of motions that the motion description made from the motion compensation vector. The 1-dimensional projection represents the motions between a pair of adjacent frames as a 1-dimensional color strip. The method connects the strips in the temporal direction and gets an image that has 1 space dimension and 1 time dimension. The resulting image has the 1-dimensional space axis and the 1-dimensional time axis as the temporal slice [16][17][18]. Our method carries information about all pixels, but the temporal slice method does only about the cross-sections. We call this image as a space-time image. Using the images, the method retrieves parts of videos as fast as image retrievals do. We can use many features defined on images.

The proposed method uses a single template space-time image, and retrieves similar plays as the play described in the template space-time image. A long space-time image template is good for retrieving the same play in the template. However, in a similar play retrieval, the long template of a space-time image is not robust about the change of the duration of a play. A short space-time image template is robust about the change of the duration in a similar play retrieval. However, a short space-time image template is weak in the discrimination power.

This paper uses both of correlation and absolute difference in similar play retrieval on only motions. First, we show the over-all structure of the proposed method. Then, space-time

image is discussed. Next, we discuss the similarity measures based on correlation and absolute difference. Then, we show the experiments on a video of a real baseball game on a Japanese TV broadcast. And last, we conclude this work.

2 Structure of the proposed method

The proposed play retrieving method is the composition of a correlation and an absolute difference of motion space-time images [6][7][8][9][10]. We describe the space-time image as ST-image in the followings. Fig. 1 shows the over-all structure of the proposed method. In fig. 1, the left side is the correlation based retrieval and the other side is the absolute difference based retrieval. In MPEG video, each 16x16 pixel block has the motion compensation vector. Using the MPEG motion compensation vector as the motion description, the size of the description is 1/64 of the original frames. In other words, the static texture description is 64 times larger than the motion descriptions in size. The static texture description is 96 times larger than the motion description in total amount with the consideration of the vector size.

The color description fits for describing the static scenes. The motion description does for describing the dynamic scenes. A play in a sport is a composition of motions. So, the motion description fits for describing plays in a sport. For retrieving the same play in a video, we can use any similarity measures. However, our goal is the retrieval of similar plays from videos. The speed of a play may change. In the case, the similarity measure must be robust about the absolute amount of motions. Our motion retrieval method uses the simple correlation as the similarity measure. The simple correlation works well in the similar play retrieval in sport videos. Using correlation as the

similarity measure, there is no hint for the absolute amount of motions. As a result, there are some error retrievals. In the relative motion space, there is no difference between large motions and small motions.

The absolute difference is a major difference feature. In similar play retrieval, the absolute difference shows poor performance in our preparation experiments. Table I shows the experimental results.

The absolute difference only shows poor performance. However, with other similarity measure the absolute difference can work some role for distinguish the similarity measured with the correlations.

The proposed similar play retrieving method uses both of the correlation and the absolute difference between a template ST-image and the ST-image from videos. We can easily differentiate the large motion difference with the absolute difference. In sport videos, the absolute difference can find the camera motions. Composing the correlation based measure and the absolute difference based measure, we construct the composite similar play retrieval method.

3 1D Degeneration from videos

There are many degeneration methods. Some works use the temporal slices [13] [19][20]. The temporal slices are easy to make and represent videos in small representation. The temporal slice is the sequence of the set of selected pixels in frames. There is no information about other pixels. The previous works make the temporal slices from color and textures in videos.

This paper makes 1D degenerated representation using the statistical features of the set of pixels. The main statistical features are mean, mode and median. This paper uses the mean for making 1D degeneration. For treating sports, the motions in videos are important.

3.1 Motion Extraction from MPEG videos and construction of space-time image

We show the process creating the space-time images of motions in an MPEG video in figure 1. From an MPEG video, we get a 2-dimensional image describing motion compensation vectors in a video. In the following, we abbreviate space-time image as ST image. The ST image has a space axis and time axis. The ST image represents the sequence of frames in one 2-dimensional image. The ST image is a very compact description of a video, and it is easy to treat.

3.2 Motion extraction from MPEG videos

First, we must have the motions in an MPEG video. An MPEG video is a sequence of GOP (Group of Pictures). Each GOP is starting from I-frame, and has B-frames and P-frames as shown in figure 2. The motion compensation vector is block-wise. The block size is 16×16 . There is a 640×480 MPEG video. The motion compensation vector image is only 40×30 pixels.

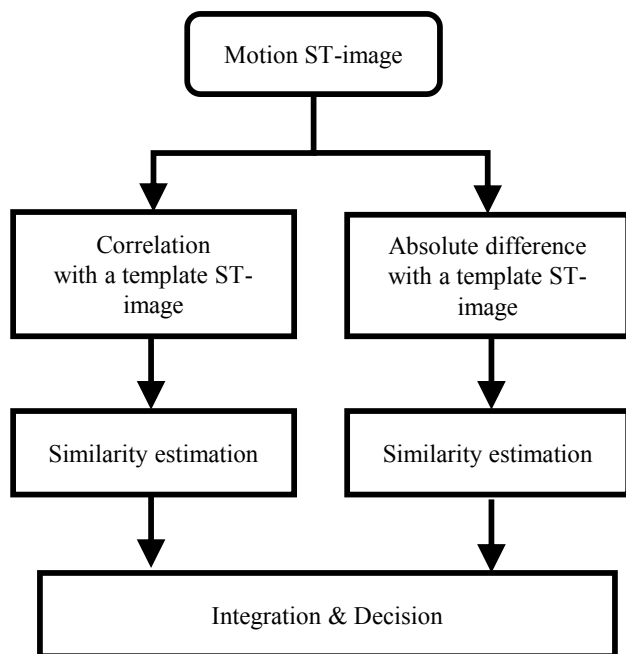


Fig. 1. Outline of the proposed similar play estimating method.

In the color frame, the grids show the motion compensation blocks. In the motion compensation vector image, the intensity of red shows the X-direction motion, and the one of green does the Y-direction motion.

MPEG videos have 3 types of frames. They are Intra-coded frame, Predicted frame and Bi-directional predicted frame. Intra-coded frame is called I-frame. Predicted frame is P-frame. Bi-directional coded frame is B-frame. The I-frame is an independent closed coded frame. There is no motion compensation vector. The P-frame has a forward prediction. The B-frame has forward and backward predictions.

There is no motion compensation vector in I-frame. The B-frame before the I-frame has a forward and a backward motion compensation vector. We use the reversed backward motion compensation vector for the motion vector of the I-frame. The P-frame has a same problem. We also use the reversed backward motion compensation vector in the B-frame just before the I-frame.

3.3 Space-time image

We have the motion vector (2-dimensional) on every motion compensation block. The amount of information is $2/16 \times 16 \times 3$ of the original color video. This is very small comparing with the original color frames. The base-ball games can long about 2 hours. This video has 200K frames. If we compare frame by frame, there needs a huge computation. There is a large difficulty to retrieve similar parts of a video.

We can retrieve similar parts of videos using classical representative frame-wise video retrieve method. However, it is difficult to retrieve similar part of videos based on the player's motions, because the motion leads a change of subsequent frames.

We can use many feature extraction methods to retrieve similar part of videos, but the applicability of the method depends on the features selected to use. The generality of the method may be lost using specified features.

This paper uses the 1-dimensional degeneration for reducing the amount of information without lost generality [3]. Fig. 2 shows the process to create a ST image from motion vector frames. We make 1-dimensional degeneration of each frame as in the top of fig. 2.

$$I_{1dx}(x) = \frac{\sum_{y \in [0, Y_{max}]} I_{2d}(x, y)}{Y_{max} + 1} \quad (1)$$

$$I_{1dy}(y) = \frac{\sum_{x \in [0, X_{max}]} I_{2d}(x, y)}{X_{max} + 1} \quad (2)$$

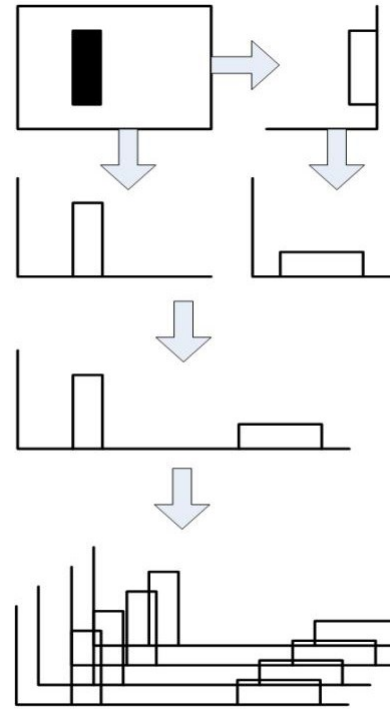


Fig. 2 Process of degrading to ST-image from the sequence of frames.

Equation (1) makes 1-dimensional degenerated description of X direction from a 2-dimensional image. Equation (2) makes 1-dimensional degenerated description of Y direction from a 2-dimensional image. In (1) and (2), $I_{2d}(x, y)$ stands for the intensity at the pixel (x, y) . (X_{max}, Y_{max}) is the coordinate of the right-upper corner.

The resulting 1-dimensional degenerated description is defined as (3).

$$I(j) = \begin{cases} j \leq X_{max} & \rightarrow I_{1dx}(j) \\ j > X_{max} & \rightarrow I_{1dy}(j - X_{max} - 1) \end{cases} \quad (3)$$

There are 2 directions to make a 1-dimensional degeneration. We use both 2 directions that are X-axis and Y-axis using (1) and (2). In each color plane, we have a 1-dimensional degenerated description. We connect the 2 degenerated descriptions onto X-axis and the transposed projection onto Y-axis as from the second to the third of figure 4. We represent the X-direction motion in red, and Y-direction motion in green. There is no value in blue. Then, we have a 1-dimensional degenerated color strip from the motion compensation vector. In the color strip, red represents the X-direction motion and green does the Y-direction motion. For the convenience, we set 255 in blue when both of X and Y direction motions are 0.

We connect the 1-dimensional color strips describing motion frames on time passing direction as the bottom of figure

4. Connecting 1-dimensional color strips, we have a color image that has 1 space axis and 1 time axis. In this paper, the image is described as ST (Space-Time) image. In the following experiments, we use the 320×240 pixels half size frames. In the MPEG format, each 16×16 pixels block holds a motion compensation vector. This leads to reduce the amount of information into $1/256$. The resulting motion image is 20×15 pixels. The 1D degenerated description is $7/60$ of the original 20×15 pixels image. As a result, the usage of the ST image of motion compensation vector in MPEG video reduces the amount of information into 0.045% from the original half size video frames. In the ST image used in this paper, X-axis holds the space and Y-axis does the time. There is no reduction of information in time axis.

The similar motion retrieval estimates what kind of motions exists on a place. It is same as the cost of the retrieval on images to retrieve similar part of videos on a ST image.

There are many similar image retrieval methods. They can be applied in ST images describing videos' motions. This paper uses the correlation between two images. We normalize the resulting correlations for compensating the variance among videos. We have 2 independent correlations between two ST images from each color plane. The blue plane exists for only the convenience for our eyes. We use an X-direction motion and a Y-direction motion in red and green planes.

Fig. 3 shows the ST-images from motions and colors. The right one is the sequence of the original video's frames. The left one is the ST-image based on the motion compensation vectors. It is enlarged in space-axis. There is no colors representing a uniform. The center one is the ST-image from the colors for comparing. In the ST-image based on colors, we can see the uniform colors.

3.4 Matching between template ST image and retrieved ST image

All ST images have same space direction size. The estimated motion image is $X \times Y$ pixels. Then, the size of the space axis of ST images is $X + Y$ pixels. For computing the

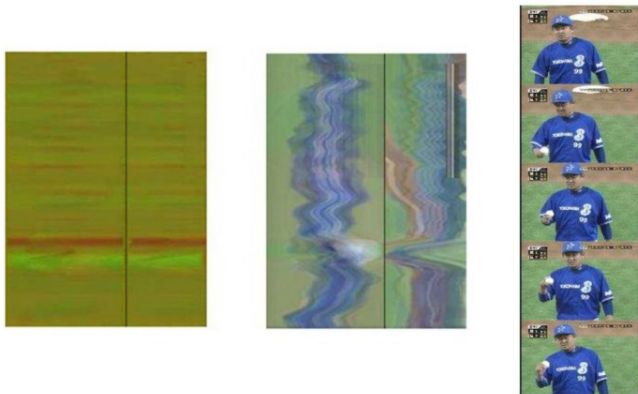


Fig. 3 Examples of ST-images.

correlations between the template ST image and any part of retrieved ST image, there is no freedom on space axis. There is only the freedom on time axis. If a template ST image is $S \times t$ and a retrieved ST image is $S \times T$, the computation cost of correlations is $S \times t \times T$.

In a baseball game, the length t of an interesting play of a video is short. So, the computational cost of correlations is small enough to be able to apply large scale video retrieval. Because of the shortness of the retrieved part, there is no need to compensate the length of the part. There is no very slow pitch or no very fast one. There is no very slow running or no very fast one.

4 Similarity measure with correlation and absolute difference in motion retrieving method

4.1 Similarity measure in motion space-time image based on correlations

We use the mutual correlation as the measure of similarity. We have 2 dimensional correlation vectors. They are X-direction motion, Y-direction motion. If there is a similar motion between the template and the retrieved part of ST image, both of the 2 correlations are large. We use the similarity measure shown in (4).

$$S_C(I_T, I_V) = \min_{p \in \{x, y\}} (NCol(I_{Tp}, I_{Vp}) - Th_p) \quad (4)$$

$$NCol(I_{Tp}, I_{Vp}) = \frac{Col(I_{Tp}, I_{Vp}) - \frac{\sum_{I_{V'p} \in V} Col(I_{Tp}, I_{V'p})}{|V|}}{SD_{I_{V'p} \in V}(Col(I_{Tp}, I_{V'p}))} \quad (5)$$

In (4), $S_C(I_T, I_V)$ is the similarity between a ST-image I_T and I_V . I_T is a template ST-image. I_V is a part of ST-image from a video. $NCol(I_{Tp}, I_{Vp})$ is the normalized correlation between I_{Tp} and I_{Vp} over a video V . The normalized correlation is normalized on the pair of the template ST-image and the ST-image from a video over the video. Th_p is the threshold. p is one of x and y that represent the X-direction motion and Y-direction motion. This similarity measure is scalar.

The equation (5) is the formal definition of $NCol$. SD is a standard derivation over the video V . Col is the correlation.

4.2 Similarity measure in motion space-time image based on absolute differences

We use the absolute difference as the measure of similarity in the motion description. In the case, we have 2 dimensional absolute differences. They are the absolute differences based on both of X-direction motion and Y-direction motion. If there is a similar motion between the template and the retrieved part of the ST image, both of the 2 absolute differences are small. We use the similarity measure shown in (6).

$$S_D(I_o, I_1) = \min_{p \in \{x, y\}} (NABD(I_{op}, I_{1p}) - Th_p) \quad (6)$$

$$NABD(I_{Tp}, I_{Vp}) = \frac{\sum_{I_{V'p} \in V} ABD(I_{Tp}, I_{V'p})}{|V|} - ABD(I_{Tp}, I_{Vp}) \quad (7)$$

$$\frac{SD_{I_{V'p} \in V}(ABD(I_{Tp}, I_{V'p}))}{SD_{I_{V'p} \in V}(ABD(I_{Tp}, I_{V'p}))}$$

In (6), $NABD(I_{op}, I_{1p})$ is the normalized negation of absolute difference between I_{op} and I_{1p} . I_{op} and I_{1p} are ST images. Th_p is the threshold. p is one of x and y that represent the X-direction motion and Y-direction motion. This similarity measure is scalar.

The equation (7) is the formal definition of $NABD$. SD is a standard derivation over the video V . ABD is the absolute difference between 2 ST-images. In (7), the terms in the numerator are placed reverse order from a normal position. This implements the negation.

5 Experiments on baseball games and evaluations

5.1 Baseball game

This paper treats baseball game MPEG videos. In baseball games, players' uniforms change between half innings. The pitch is the most frequent play in a base-ball game. There is large number of pitches. This paper uses a single play of a pitch as a template. Using this template, the proposed method retrieves large number of pitches using similar motion retrieval.

Motion based similar video retrieval can find many types of plays based on the template. There are a few repeated plays that are not pitches. This paper distinguishes a pitch and other plays.

5.2 Experimental objects

This paper uses a whole base-ball game for experiments. The game is 79minutes, 132485 frames in a video. In the game,

there are right-hand pitchers and a left-hand pitcher. There are 168 pitches. There are 31648 frames that represent the camera work that catches the pitching scenes.

Fig. 4 shows the example of pitches in our experimental video at each 5 frames distance. The center one and right one differ from the left one at the uniforms. The right one differs from the left one at left-hand and right-hand.

5.3 Experiment process

The experimental videos are recorded from Japanese analog TV to DVD. Then, the recorded videos are reduced into 320×240 pixels and encoded MPEG1 format. Most plays of pitches are very short. So there is no reduction on time direction. There are 30 frames in 1S. There are all parts including telops, sportscasters and CG overlays. The first step of our experiment is the extraction of motion compensation vectors. The motion compensation vector is at each 16×16 pixel blocks. In every motion compensation block, we have a motion compensation vectors. The similar play retrieval in motion frames uses 20 frames of the pitch as in fig. 2 as a template and retrieves the shots including pitches of a video. In Fig. 3, the left one shows the part of the ST-image based on the motions from motion compensation vectors in a MPEG video. And, the center shows the ST-image based on the original colors in frames. The frames of this part are shown in the right.



Fig. 4 Examples of pitching shots.

In the color-based ST-image, we can see the place of the player. In the motion-based Space-time map, we can see the stripes representing the player's motion.

5.4 Correlation based similarity measure in pitching retrieval

These experiments use a single template image of the length 20 frames. When we say the template that starts 83000th frame, we use 83000th frame to 83020th frame. There are 20 motions. This sequence has 21 frames. Our pre-experiments using some length of template ST-image show that the 20 motion frames template ST-image is best. Here after, we use the template of the length 20. We control the thresholds that make the F-measure as the maximum using Excel goal seek function.

In this experiment, we use 3 template pitches. They are the template starts from 120130th, 123340th and 191360th. The F-measures spans from 0.503 to 0.807. Table I shows the experimental results. We try some templates. The template started from 120130th frame is best. We select these 3 templates that includes the best one.

In number, in the best case, the method finds 131 pitches in 168 pitches. The method retrieves 157 candidate pitches in 132485 possible candidates. This means that the 26 error retrievals in 132317 no candidates. This is 99.98% precision. In this case, there is huge amount of frames. There is huge unbalance between yes samples and no samples. In the case, it is difficult to get a high F-measure.

5.5 Absolute difference based similarity measure in pitching retrieval

In this experiment, we use the same 3 template pitches in the correlation based method. They are the template starts from 120130th, 123340th and 191360th. The F-measures spans from 0.315 to 0.371. The performance is very low. Table II shows the experimental results. In number, the method finds 165 pitches in 168 pitches in the video. The method retrieves 708 error candidates. This shows that the absolute difference based similarity cannot distinguish a pitch from other motions.

5.6 Combination both of correlations and absolute differences

For retrieving the precise pitches in frames, we need to use the correlation based similarity. With combining the correlation based similarity and the absolute difference based similarity, we can improve the performance of the precise pitch retrieval.

Using the correlation based similarity or the absolute difference based similarity, we have no difficulties to find the proper set of thresholds. In the correlation based similarity or the absolute difference based similarity, there are 2 thresholds that work in each X and Y direction motions.

To combine the correlation based similarity and the absolute difference based similarity, we use logical

TABLE I
Performance with correlation

Template	Recall	Precision	F-measure
120130	78.0%	83.7%	0.807
123340	70.8%	77.8%	0.742
191360	58.9%	43.9%	0.503

TABLE II
Performance with absolute difference

Template	Recall	Precision	F-measure
120130	98.2%	18.9%	0.371
123340	100.0%	18.7%	0.315
191360	100.0%	18.8%	0.316

conjunction. We can use logical disjunction. However, minimum is extension of logical conjunction. We select logical conjunction in this paper. We have 4 thresholds in the combination. It is difficult to optimize all thresholds at once with short processing time. We divided the optimization of thresholds into 2 steps. They are the correlation based similarity threshold optimization and the absolute difference based one.

First, we optimize the thresholds in the absolute difference based similarity. Then, we do ones in the correlation based similarity. In the absolute difference based experiment, the recalls are high enough. This leads the decision. Of cause, there is no difference between the correlation based first and the absolute difference based first.

Table III shows the result of the experiments. The templates are same in former experiments. The resulting F-measures spans from 0.745 to 0.865. In every template, the performance increases much. Especially, in the case of the template starting 191360th, there is 48% increase in F-measure.

6 Conclusions

This paper discusses about the retrieval of similar plays in sport MPEG videos using similar motion retrieval based on both of correlation and absolute difference. For recognizing sport videos, the motions represent important meanings. In the cases, there must be similar video retrieval methods based on the motions described in the videos. The proposed similar play retrieval method is the combination of correlation and absolute difference motion based on only motion compensation vectors in MPEG videos.

TABLE III.
Performance with the combination of correlation and absolute difference

Template	Recall	Precision	F-measure
120130	94.0%	80.0%	0.865
123340	83.9%	87.9%	0.859
191360	84.5%	66.6%	0.745

The experiment shows that the similar play retrieval works well using both of correlation and absolute difference on ST-images made from motion compensation vectors in MPEG videos. Classical works using MPEG motion compensation vector only uses global-scale motions. However, the proposed method utilizes local motions. The proposed combination of correlation based similarity measure and absolute difference based similarity measure works well in our experiments. Using both similarity measures, the proposed method gets some more performance than a single correlation based similarity measure base on motion based ST-images.

REFERENCES

- [1] Hisashi Miyamori, Eiji Kasutani, Hideyoshi Tominaga, "Video Retrieval Method by Query Using Action Phrases", Trans. JIEICE, 80-D-2, 6, pp.1590-1599, 1997.
- [2] Akira Kamegaya, Hirotsugu Kinoshita, "An image retrieving method using the object index and the motion", JIEICE technical report. IE, 23, 8, pp.43-48, 1999.
- [3] Tomohiro Matsuike, Ryoji Ishii, Shinichiro Maeda, Yoshihiro Okada, "Spatio-Temporal Image Processing for Retrieving Similar Video Sequences", JIEICE technical report. PRMU, 107, 427, pp.299-304, 2007.
- [4] Akio Nagasaka, Takafumi Miyatake, Hirotada Ueda, "Realtime Video Scene Detection based on Shot Sequence Encoding", Trans. JIEICE, 79-D-2, 4, pp.531-537, 1996.
- [5] Yoshitomo Yaginuma, Motofumi Suzuki, Yasutaka Shimizu, "Retrieval of Educational Image Contents Based on Color Features and Keywords", JIEICE technical report. ET, 106, 507, pp.111-116, 2007.
- [6] Kyota Aoki, "Video retrieval based on motions", JIEICE technical report, IE, 108, 217, pp. 45-50, 2008.
- [7] Kyota Aoki, "Plays from Motions for Baseball Video Retrieval", Computer Engineering and Applications, 2010 Second International Conference on, pp.271-275, 2010.
- [8] Kyota Aoki, "Play estimation using multiple 1D degenerated descriptions of MPEG motion compensation vectors", Computer Sciences and Convergence Information Technology, 2010 5th International Conference on, pp. 176-181, 2010.
- [9] Kyota Aoki, "PLAY ESTIMATION USING SEQUENCES OF MULTIPLE 1D DEGENERATED DESCRIPTIONS OF MPEG MOTION COMPENSATION VECTORS", Proceedings of the Eighth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications, pp. 268-275, 2011.
- [10] Kyota Aoki, Takuro Fukiba, "Play Estimation with Motions and Textures in Space-Time Map Description", Workshop on Developer-Centred Computer Vision in ACCV 2012, LNCS 7728, pp.279-290, 2012.
- [11] Michael Fleischman, Deb Roy, "Situated Models of Meaning for Sports Video Retrieval", Proc. NAACLHLT2009, Companion Volume, pp. 37-40, 2007.
- [12] N. Sebe, M.S. Lew, X. Zhou, T.S. Huang, E.M. Bakker, "The State of the Art in Image and Video Retrieval", LNCS2728, pp7-12, 2003.
- [13] S.-L. Peng, "Temporal slice analysis of image sequences", Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91. IEEE Computer Society Conference on, pp. 283-288, 1991.
- [14] Kyota Aoki, "Block-wise High-speed Reliable Motion Estimation Method Applicable in Large Motion", JIEICE technical report. IE, 106, 536, pp.95-100, 2007.
- [15] Masashi Nobe, Yuji Ino, Kyota Aoki, "Pixel-wise Motion Estimation Based on Multiple Motion Estimations in Consideration of Smooth Regions", JIEICE technical report. IE, 106, 423, pp.7-12, 2006.
- [16] Chong-Wah Ngo, Ting-Chuen Pong, and Hong-Jiang Zhang, "On Clustering and Retrieval of Video Shots Through Temporal Slices Analysis", IEEE Trans. on Multimedia, 4, 4, pp. 446-458, 2002.
- [17] Patrick Bouthemy, Ronan Fablet, "Motion Characterization from Temporal Cooccurrences of Local Motion-based Measures for Video Indexing", 14th Int. Conf. on Pattern Recognition, ICPR'98, Brisbane, pp. 905-908, August 1998.
- [18] E Keogh, CA Ratanamahatana, "Exact indexing of dynamic time warping", Knowledge and Information Systems, 7, 3, pp.358-386, 2005.
- [19] C.W. Ngo, T.C. Pong, R.T. Chin, "Detection of gradual transitions through temporal slice analysis", Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., 1, pp.41, 1999.
- [20] Chong-Wah Ngo, Ting-Chuen Pong, Chin, R.T., "Video partitioning by temporal slice coherency", Circuits and Systems for Video Technology, IEEE Transactions on, 11, 8, pp. 941-953, 2001.

SESSION

**IMAGE AND SIGNAL PROCESSING
APPLICATIONS AND NOVEL ALGORITHMS**

Chair(s)

TBA

Decision-Level Fusion of Multi-modal Data using Manifolds

Yassine Belkhouche
Emerging Analytics Center
University of Arkansas at Little Rock
mybelkhouche@ualr.edu

Bill Buckles
Department of computer science and engineering
University of North Texas
bbuckles@cse.unt.edu

Abstract—The advancement in sensing technology created a need for developing data fusion techniques. In This paper we address the problem of fusing 3D LiDAR data with visual imagery. The purpose of our fusion scheme is the classification of LiDAR data into four different classes. We establish a decision-level fusion scheme to solve this problem using manifolds. Our method proceeds in three steps. First, We used features extracted from each modality to learn a separate manifold. In the second step, we defined a classification confidence level for each class on each manifold using training data. Finally, in order to predict the class of new data point, we predict the class of that point on each manifold separately. Using the classification confidence, We established a decision-level scheme that combines the individual prediction on each manifold into a final prediction. We test our method using two data sets. Results show the effectiveness of our approach for decision-level fusion of multi-modal data.

I. INTRODUCTION

Multi-modal data fusion is very important for many civilian and military applications. In this paper, we consider decision-level fusion of registered LiDAR and visual imagery [1], [2], [3], [4]. We introduced a manifold-based fusion technique. The proposed method starts by learning two different manifolds. The first manifold is learned using LiDAR features, while the second manifold is learned using visual features. Using a decision made on each manifold individually, we proposed a decision-making scheme leading to more accurate decisions. This paper is organized as follows. The fusion problem is addressed in section II. A review on fusion and manifold learning techniques is provided in section III. The manifold-based fusion is introduced in section IV. A study case is provided in section V to illustrate the proposed fusion technique. We conclude the paper in section VI.

II. PROBLEM DESCRIPTION

Given two modalities MD_1 and MD_2 , each data point in these modalities belong to one of the classes $\{c_1, \dots, c_k\}$. Let $\{c_1, \dots, c_i\}$ be the classes recognizable using MD_1 and $\{c_j, \dots, c_k\}$, be the classes recognizable using MD_2 . The objective is to develop a fusion method that combines information from MD_1 and MD_2 in order to maximize the number of recognizable classes and reduce the miss classification errors. This problem is illustrated in figure 1.

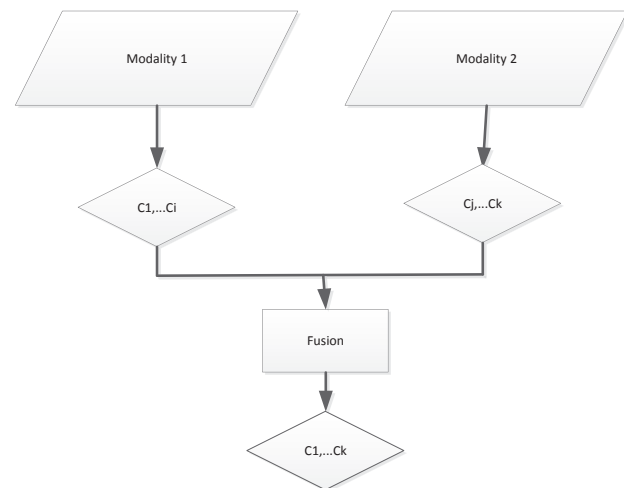


Fig. 1. Multi-modal fusion problem.

III. REVIEW ON MANIFOLD LEARNING AND DATA FUSION TECHNIQUES

Several well-developed theories have been used for decision-level data fusion, among these are: The Bayesian inference [5], [6], Dempster-Shafer evidential theory [7], neural network, voting logic and machine learning and classification algorithms.

Hugh [8] classified the objectives of data fusion into three main categories:

- **Complementary:** The datasets to be fused are independent of each other, but they can be combined to provide more complete information.
- **Competitive:** When the measurements obtained for the same property from different sensors are independent, the objective of fusion is to determine which measurement is better.
- **Cooperative:** In this situation, the measurements obtained from different sensors are combined to extract information that are not available from individual sensors.

Dasarathy [9] introduced a fusion model based on the nature of the input and the output properties of the fusion model. He identify five different possibilities: Data-In/Data-Out, Data-In/Features-Out, Data-In/Decision-Out, Features-In/Features-Out, Features-In/Decision-Out and Decision-In/Decision-Out.

Manifold learning is new concept introduced to approach machine learning algorithms such as data clustering, pattern classification, interpolation, detection, compression, and several other applications. Manifold learning algorithms assume that high dimensional data points lay near or on a low dimensional manifold that has to be learned from the input data, therefore manifold learning can be perceived as dimensionality reduction technique. Several algorithm have been introduced to learn the topology of the underlying data such as Isomap, MDS, LLC and many others.

Dimensionality reduction methods are classified into two main classes: linear and nonlinear. Linear methods such as principal components analysis (PCA) [10], linear discriminant analysis (LDA) have success when the dataset is linearly distributed. However real world data are highly nonlinear, therefore linear techniques are not suitable. In the rest of this section, we review the most important algorithms used for non-linear dimensionality reduction.

Multidimensional scaling (MDS) [11] was initially introduced for visualization purposes. MDS reduce the dimensionality of the data by preserving the Euclidean distance between data samples in the original space and their corresponding embedding in the low dimensional space. Distances from each point to all others are stored in a distance matrix that represents the geometric properties of the original data. The projected data is generated in order to minimize the mean square error between the original distance matrix and that of the projected data set.

The isometric feature mapping (Isomap) algorithm was introduced by [12]. This algorithm recovers the manifold structure by preserving geodesic distances between all points in the dataset. Three main steps are considered to learn a manifold using Isomap. The algorithm starts by constructing the k-nearest neighbors graph. Using this graph, Isomap computes the geodesic distances between all pairs of data points. In the last step, a low dimensional embedding is generated using the geodesic distances as an input to the MDS algorithm.

Laplacian Eigenmap (LE) introduced by [13], is another method for manifold learning. That method reconstructs the manifold structure from high dimensional data by preserving locality, which emphasizes the natural clusters in the data. LE proceeds in three steps. In the first step, the adjacency graph is build using k-nearest neighbors or the ϵ -neighbors. The second step is weights assignment. Each edge in the graph is given a weight. Two ways have been proposed for weights assignment: One way is to use the heat kernel, the second way is to assign 1 to entries in the adjacency matrix where there are connections and 0 otherwise. In the third step, an eigenvector problem is formulated. Given the solution to that problem only a number of eigenvectors that are smaller than the original dimensionality of the data are considered to find a low dimensional embedding.

Locally Linear Embedding (LLE), have been proposed in

[14]. This method attempts to preserve the local neighborhood structure of the data, it consists of three steps. The first step is identifying the k-nearest neighbors for each data point. In the second step, weights are computed by minimizing a reconstruction error cost function. These weights are used in the final step to compute the best low dimensional embedding.

Other works that use manifold for data fusion were proposed in [15], [16], [17], [18].

IV. DECISION LEVEL FUSION SCHEME USING MANIFOLDS

In this section, we propose a decision-level fusion scheme to solve the problem described in section II. Given two modalities, MD_1 and MD_2 and a set of features Fv_1 characterizing data points in MD_1 and features Fv_2 characterizing data points from MD_2 . Let $\{c_1, c_2, \dots, c_k\}$ be a set of classes. Each data points belong to some class c_i . The geodesic distance between the centroid point p of the data points in the class c_i and the centroid point q of the data points in the class c_j on a manifold \mathcal{M} is defined as follows:

$$d_{\mathcal{M}}(p, q) = \inf\{L(\gamma) : \gamma(0) = p, \gamma(1) = q\}. \quad (1)$$

where $\gamma : [0, 1] \rightarrow \mathcal{M}$, is the curve joining the points p and q . $L(\gamma)$ is the length of the curve γ . The projection of a point r on the curve γ is the point s that minimize the geodesic distance between the point r and a point s on the curve γ . It is defined using equation 1, with an additional constraint that $s \in \gamma$. Let \mathcal{M}_1 be the manifold learned using the feature set Fv_1 and \mathcal{M}_2 be the manifold learned from the feature set Fv_2 . Let p and q be the centroids of the classes c_i and c_j respectively. All the points in the classes c_i and c_j are projected on the curve γ joining p and q . Let r be the point that maximizes the geodesic distance between the point p and all the projected points from class c_i on the curve γ . Similarly, let s be the point that maximizes the geodesic distance between the point q and all the projected points from class c_j on the curve γ . The classification confidence on the manifold \mathcal{M} associated with classes c_i and c_j is computed as follows:

$$S_{c_i, c_j} = 1 - \frac{d_{\mathcal{M}}(p, r) + d_{\mathcal{M}}(q, s)}{d_{\mathcal{M}}(p, q)}. \quad (2)$$

where S_{c_i, c_j} indicate the confidence of correctly classifying the classes c_i and c_j on the manifold \mathcal{M} . A value closer to one indicates a high confidence, while a value closer to zero indicates a low confidence. The classification confidence is computed for every pair of classes separately for both manifolds \mathcal{M}_1 and \mathcal{M}_2 . This indicator is computed using a training set. In the testing phase, each data point is assigned a label on each manifold using the geodesic distance:

$$\mathcal{M}(c_i) = \operatorname{argmin}(d_{\mathcal{M}}(x_i - x_j)). \quad (3)$$

where x_i is a testing point and x_j is a training point, and $\mathcal{M}(c_i)$ is the label assigned to x_i on manifold \mathcal{M} . Let $C_{M_i} = M_i(S_{M_i(c_k), M_j(c_k)})$ be the classification confidence associated with the classes $M_i(c_k)$ and $M_j(c_k)$. Given the label assigned to a testing point on manifolds \mathcal{M}_1 and \mathcal{M}_2 , we propose the

following decision level fusion technique:

$$c_i = \begin{cases} \mathcal{M}_1(c_i) & \text{if } \mathcal{M}_1(c_i) = \mathcal{M}_2(c_i) \\ \mathcal{M}_1(c_i) & \text{if } \mathcal{M}_1(c_i) \neq \mathcal{M}_2(c_i) \text{ \& } \mathcal{C}_{M_1} > \mathcal{C}_{M_2} \\ \mathcal{M}_2(c_i) & \text{if } \mathcal{M}_1(c_i) \neq \mathcal{M}_2(c_i) \text{ \& } \mathcal{C}_{M_1} < \mathcal{C}_{M_2} \end{cases} \quad (4)$$

where c_i is the final classification of the instance x_i .

V. STUDY CASE

We used the proposed fusion scheme to classify data segments from a registered LiDAR and visual imagery. We consider four classes: Buildings (B), trees (T), grass (G), and roads (R). Using LiDAR data, one can easily distinguish between buildings and trees, however it is very difficult to distinguish grass from road segments. Grass and road segments, can be easily distinguished using visual features. Trees and grass cannot be distinguished using visual features. This section consists of two parts. The first part is dedicated to features extraction from LiDAR and visual imagery. In the second part, we discuss test cases used to validate the proposed technique.

A. Features Extraction From LiDAR

In this section we present a set of features used to characterize, classify, or cluster LiDAR data. Two types of feature are considered: point-based features and segment-based features. Many of these features have been introduced and used in the literature [19], [20].

1) *Flatness*: In this section, we describe two ways to compute the flatness of a 3D area represented by a set of points. The first method to compute the flatness value, is to divide the area surface in 2D by its corresponding volumetric surface in 3D. This can be achieved by using triangulated irregular network (TIN) representation of the points. The 2D surface will be the sum of triangles area in 2D (ignoring the z value), while the 3D area is the sum of the triangles area in 3D. A flat area will have a flatness value closer to one, while a non-flat area will have a value closer to zero. The flatness computed by this method is described in equation (5):

$$f = \frac{2D \text{ area}}{3D \text{ area}}. \quad (5)$$

The second method to compute the flatness value, is based on the k -nearest neighbors graph. Given a point for which we want to compute the flatness value. We start by finding the k -nearest neighbors of this point, then the flatness is computed as the sum of the graph edges in 2D divided by its corresponding sum in 3D:

$$f = \frac{\sum_{i=1}^k e_{i_{2D}}}{\sum_{i=1}^k e_{i_{3D}}}. \quad (6)$$

where e_i is the length of the edge i .

2) *Normal variation*: The normal variation is computed as follows. Given a point p , let $N(p)$ be the k -nearest neighbors of p . A TIN interpolation is constructed using the points in the set $N(p)$. A normal vector is computed for each triangle in this representation. Each normal vector have three components: n_x ,

n_y and n_z . Using these components, we compute the following measures:

- The mean of normals with respect to x , y and z components ($mean(n_{x_i})$, $mean(n_{y_i})$, $mean(n_{z_i})$).
- The standard deviation of normals with respect to x , y and z components ($std(n_{x_i})$, $std(n_{y_i})$, $std(n_{z_i})$).

3) *Elevation difference*: Given the k -nearest neighbors graph, we suggest four different measures of the elevation difference. The maximum elevation difference, witch consists of the maximum elevation difference between a point of interest to each of its neighboring points. Similarly the minimum elevation difference is the minimum of all elevation differences from a point to all its neighbors. The average elevation difference witch is the mean of all the elevation differences from one point to all its neighbors. The forth measure of elevation difference is the standard deviation of elevation differences between a point an all its neighbors. Given a point p_i and its k -nearest neighbors points $N(p_i)$, equations (7)-(10) shows the computation of max, min and mean, and standard deviation elevation difference measures, respectively.

$$MaxEl_{p_i} = \max(abs(z_{p_i} - z_{p_j})) \quad p_j \in N(p_i) \quad (7)$$

$$MinEl_{p_i} = \min(abs(z_{p_i} - z_{p_j})) \quad p_j \in N(p_i) \quad (8)$$

$$AvgEl_{p_i} = \frac{1}{|N(p_i)|} \sum_{j=1}^{|N(p_i)|} abs(z_{p_i} - z_{p_j}) \quad p_j \in N(p_i) \quad (9)$$

$$StdEl_{p_i} = \sqrt{\frac{1}{|N(p_i)|} \sum_{j=1}^{|N(p_i)|} (z_{p_i} - \mu_z)^2} \quad p_j \in N(p_i) \quad (10)$$

The feature vector associated with each LiDAR point p_i is the following:

$$v_i = [z \ f \ Mean.n_x \ Mean.n_y \ Mean.n_z \ Std.n_x \ Std.n_y \ Std.n_z \ MaxEl \ MinEl \ AvgEl \ StdEl]$$

The set of features discussed above are point-based features. In the rest of this section, we discuss segment-based features. Given a segment that consists of n points, we characterize this segment using the following features:

- *Segment Flatness*: Similarly to point-based flatness, segment-based flatness can be computed in two ways. The first way is to build the k nearest neighbors graph using the points in the segment, and then compute the flatness of the segment as the sum of the edge lengths in 2D divided by their corresponding sum in 3D. The second way is based on Delaunay triangulation. In the second method, the flatness is computed as the sum of the triangles surface in 2D, divided by the sum of the triangles surface in 3D.
- The mean of the x -component of the normals.
- The mean of the y -component of the normals.
- The mean of the z -component of the normals.
- The standard deviation of the x -component of the normals.
- The standard deviation of the y -component of the normals.

- The standard deviation of the z-component of the normals.
- Segment maximum elevation difference: Is the maximum elevation difference between any two points in the segment.
- Segment minimum elevation difference: Is the minimum elevation difference between any two points in the segment.
- Segment standard deviation of the elevation difference: This measures the standard deviation of the elevation differences between points in the segment.

B. Feature Extraction From Visual imagery

To classify segments from visual image, we used pixels values in the RGB and the HSV domains. Each segment is represented using a feature vector consisting of the average R value, average G value, average B value, average H value, average S value, and average V value.

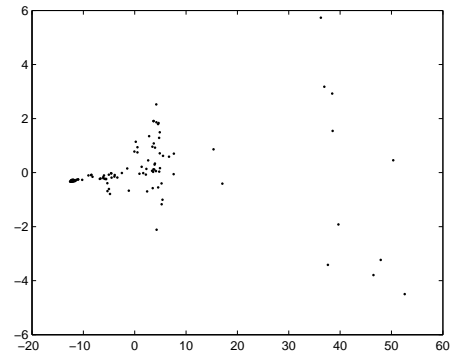
C. Tests

We used two datasets. The first area is $750\text{ m} \times 410\text{ m}$, it consists of 35188 points. The second area is $750\text{ m} \times 400\text{ m}$, and consists of 34876 points. We manually labeled a set of objects on both LiDAR and its corresponding visual image. Table I summarizes the characteristics of each area. We used Isomap to learn two manifolds, M_l using LiDAR features, and M_v using visual features. Figure 2 shows the learned manifolds from LiDAR and visual features respectively (each point in this figure is a segment representing one of the four classes). For each dataset, 70% of samples were randomly selected and used for learning, while the other 30% of samples were used for testing. We computed the accuracy of correctly classified samples, as well as the confusion-matrix. Tables II and III show the confusion matrix for area 1, using two different testing subsets. Using the first testing subset, there are four grass segments classified as road segments using the LiDAR manifold, and one road segment classified as building using the visual manifold. When combining the classification results from the two manifolds using our fusion scheme, all segments are correctly classified. Using the second testing subset, there are five grass segments classified as road segments using the LiDAR manifold, one road segment classified as building and three buildings classified as road segments using the visual manifold. When combining the classification results from the two manifolds using our fusion scheme, all segments are correctly classified.

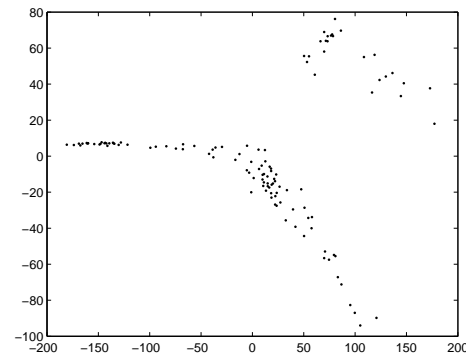
	Number of segments	Buildings	Trees	Grass	Roads
Area 1	121	67	10	15	29
Area 2	64	20	14	10	20

TABLE I
CHARACTERISTICS OF THE TEST CITES.

Tables IV and V show the confusion matrix for area 2, using two different testing subsets. These results show that the proposed fusion method performs better than the individual decision derived from each manifold. Using the first



(a) Manifold learned using LiDAR features



(b) Manifold learned using visual features

Fig. 2. Manifolds learned from LiDAR and visual features for area 1.

testing subset, there are two grass segments classified as road segments and one road segment classified as tree using the LiDAR manifold. Using the visual manifold, there is one road segment classified as building and one grass segment classified as tree. When combining the classification results from the two manifolds using our fusion scheme, there is only one road segment miss classified. Using the second testing subset, there are two grass segments classified as road segments and one road classified as tree using the LiDAR manifold. Using the visual manifold, there is one building segment classified as road and one road classified as building. When combining the classification results from the two manifolds using our fusion scheme, there is only one road miss classified as tree.

VI. CONCLUSION

In this paper we proposed a new idea for decision-level fusion of 3D LiDAR data and visual imagery. The purpose of our fusion method is the classification of LiDAR data into four different classes. We used manifolds learned from features extracted from each modality, and we established a decision-level method for fusing these modalities. We defined a confidence level associated with the classification decision made on each manifold. This confidence level is used in our fusion scheme to select the best classification. A case study was provided to illustrate the effectiveness of our method.

		Classified			
		B	T	G	R
B	18	0	0	0	0
T	0	3	0	0	0
G	0	0	1	4	
R	0	0	3	3	

		Classified			
		B	T	G	R
B	18	0	0	0	0
T	0	3	0	0	0
G	0	0	4	0	
R	0	0	0	7	

TABLE II

DATASET 1. (A): LiDAR CLASSIFICATION. (B): VISUAL CLASSIFICATION.
(C): FUSED CLASSIFICATION.

		Classified			
		B	T	G	R
B	19	0	0	0	0
T	0	2	0	0	0
G	0	0	0	5	
R	0	0	5	4	

		Classified			
		B	T	G	R
B	19	0	0	0	0
T	0	2	0	0	0
G	0	0	5	1	
R	0	0	0	8	

TABLE III

DATASET 1. (A): LiDAR CLASSIFICATION. (B): VISUAL CLASSIFICATION.
(C): FUSED CLASSIFICATION.

		Classified			
		B	T	G	R
B	17	0	0	0	0
T	0	3	0	0	0
G	0	0	4	0	
R	1	0	0	7	

		Classified			
		B	T	G	R
B	4	0	0	0	0
T	0	2	0	0	0
G	0	0	1	2	
R	0	1	1	3	

TABLE IV

DATASET 2. (A): LiDAR CLASSIFICATION. (B): VISUAL CLASSIFICATION.
(C): FUSED CLASSIFICATION.

		Classified			
		B	T	G	R
B	6	0	0	0	0
T	0	3	0	0	0
G	0	0	0	2	
R	0	1	2	2	

		Classified			
		B	T	G	R
B	6	0	0	0	0
T	0	3	0	0	0
G	0	0	2	0	
R	0	1	0	4	

TABLE V

DATASET 2. (A): LiDAR CLASSIFICATION. (B): VISUAL CLASSIFICATION.
(C): FUSED CLASSIFICATION.

		Classified			
		B	T	G	R
B	3	0	0	0	0
T	0	2	0	0	0
G	0	1	2	0	
R	1	0	0	5	

		Classified			
		B	T	G	R
B	5	0	0	1	
T	0	4	0	0	
G	0	0	2	0	
R	1	0	0	3	

REFERENCES

- Min Ding and Avidesh Zakhor, "Automatic registration of aerial imagery with untextured 3d LiDAR models," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, June 2008, pp. 1–8.
- A. Mastin, J. Kepner, and J. Fisher, "Automatic registration of lidar and optical images of urban scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, June 2009, pp. 2639–2646.
- Tae-Suk Kwak, YongIl Kim, Ki-Yun Yu, and Byoung-Kil Lee, "Registration of aerial imagery and aerial LiDAR data using centroids of plane roof surfaces as control information," *KSCIE journal of Civil Engineering*, vol. 10, no. 5, pp. 365–370, 2004.
- Ayman Habib, Mwafag Ghanma, Michel Morgan, and Rami Al-Ruzzouq, "Photogrammetric and LiDAR data registration using linear features," *Photogrammetric Engineering and Remote Sensing*, vol. 7, no. 6, pp. 699–707, 2005.
- Michael E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," *Advanced Lectures on Machine Learning*, pp. 41–62, 2004.
- Biao Chen and Pramod K. Varshney, "A bayesian sampling approach to decision fusion using hierarchical models," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1809–1818, 2002.
- S. Le Hégarat-Masclé, D. Richard, and C. Ottlé, "Multi-scale data fusion using dempster-shafer evidence theory," *Integr. Comput.-Aided Eng.*, vol. 10, no. 1, pp. 9–22, Jan. 2003.
- H. F. Durrant-Whyte, "Sensor models and multi-sensor integration," *The International Journal of Robotics Research*, vol. 7, no. 6, pp. 97–113, 1988.
- B. V. Dasarathy, *Decision Fusion*, IEEE Society Press, 1 edition, 1994.
- G. H. Dunteman, *Principal components analysis*, Sage Publications, 1989.
- I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications*, Springer-Verlag New York, 2nd edition, 2005.
- J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.
- M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2002.
- L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of nonlinear manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.
- M. A. Davenport, C. Hegde, M. F. Duarte, and R. G. Baraniuk, "Joint manifolds for data fusion," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2580–2594, 2010.
- A. A. Jamshidi, M. J. Kirby, and D. S. Broomhead, "Geometric manifold learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, March 2011.
- Yoshua Bengio, Jean Francois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering," in *In Advances in Neural Information Processing Systems*. 2004, pp. 177–184, MIT Press.
- M. Yang, "Extended isomap for pattern classification," in *Proceedings 16th International Conference on Pattern Recognition*, CA, USA, 2002, pp. 615–618.
- J. Secord and A. Zakhor, "Tree detection in aerial LiDAR and image data," Atlanta, GA, October 2006, pp. 2317–2320.
- A.P. Charaniya, R. Manduchi, and S.K. Lodha, "Supervised parametric classification of aerial LiDAR data," June 2004.

DT-Binarize: A Hybrid Binarization Method using Decision Tree for Protein Crystallization Images

İmren Dinç¹, Semih Dinç¹, Madhav Sigdel¹, Madhu S. Sigdel¹, Marc L. Pusey², Ramazan S. Aygün¹

¹DataMedia Research Lab, Computer Science Department
University of Alabama in Huntsville
Huntsville, Alabama 35899

²iXpressGenes Inc., 601 Genome Way, Huntsville AL, 35806

Email: {id0002, sd0016, ms0023, mss0025, aygunr} @uah.edu¹, marc.pusey@ixpressgenes.com²

Abstract—A single thresholding technique may not provide the best binarization for all images of datasets such as protein crystallization images. To overcome this limitation, multiple thresholding methods are used to binarize images. Whenever multiple thresholding techniques are used, it is important to know which one provides the best result automatically. To solve this problem, in this study, we propose an alternative technique for image thresholding that employs a tree based structure to determine the best thresholding approach for a particular case. The leaf nodes of the tree indicate different global thresholding techniques, which have different abilities to binarize the image. We try to select the best approach by making decisions that are based on the characteristic features of the sample such as standard deviation.

We have applied this technique to our protein image dataset and compared the results with the ground truth binary images that are manually generated by experts. Experimental results indicate that using a selecting the best one in a group of global thresholding methods is beneficial rather than single one. We provide the comparison results using some well-known accuracy measures. Our technique has reached 0.82 using Matthew's correlation coefficient (MCC) and increased the MCC value by 0.11.

Keywords: Global Thresholding, Image Binarization, Protein Crystallization, Decision Tree

1. Introduction

Protein crystallography is an important research area that allow scientists to study structure of the proteins. The structure gives information about the protein functionality, which is one of important steps of drug discovery in medicine [1]. Protein crystallization process is a complex task that comprises of several stages. Every stage requires high attention since some parameters, such as pH and temperature, need to be set carefully in order to grow protein crystals. In addition, growing a crystal usually requires many trials and most of the trials does not yield a desired protein crystal [2].

In non-automated systems, scientists check hundreds of images of protein samples to find the crystal form. Since a

protein crystal rarely occurs, it may take very long time to detect desired samples by checking manually [1]. For this reason, detecting and classifying protein crystals using an automated system is significantly important for the scientists to save time and effort. Automated systems typically use geometrical features of the protein image such as lines, shapes, area, and perimeter to distinguish crystals. Before extracting these features, a binarization (or thresholding) stage that is very critical to extract reliable geometrical features is required.

Image binarization is not a simple task and there is not an optimal solution that works for all cases. In the literature, there are many studies that focus on different aspects of the problem as global, local, or adaptive thresholding. Studies focus on their own problem domain to find the best approach for binarization [3].

Usually, crystal images are expected to have distinguishable features such as high intensity, sharp clear edges and proper geometric shapes. However, in some cases these features may not be dominant due to focusing or reflection problems even if there is a protein crystal in the image. For that reason, a single type of thresholding technique may not provide an informative binary image to use in classification of the images. Moreover, binarized image may lose some important information or it may keep some unnecessary information. This may yield incorrect classification. For example, incorrect thresholding method may cause to lose a blurred crystal in the image.

In our previous work [4], we used three thresholding techniques (Otsu's Threshold, 90th Percentile Green Intensity Threshold, Max Green Intensity Threshold) together to classify protein crystallization images not to lose any informative feature. However, we noticed that we also have included unnecessary features which may cause incorrect classification results. To avoid this problem, in this study, we propose an alternative approach that selects the best thresholding technique for a particular image using decision trees. Using some statistical features of the images we train a decision tree using pre-labeled samples. Leaf nodes of the tree indicates a thresholding technique that properly fits for that particular case. In the test stage, using the same statis-

tical features of the test sample, we decide the thresholding method that provides best results. Our technique selects the most informative and reliable binary image of the protein crystal. In this way, the complexity of our system may be reduced since we are dealing with less number of features (i.e., features from a single thresholded image are used rather than from multiple thresholded images). Our method is a hybrid method since it uses multiple thresholding techniques. Since our method used decision trees we call our method as *DT-Binarize*.

This research uses protein crystallization images dataset provided by iXpressGenes, Inc. As our earlier work, we classify the protein images into three main groups (noncrystals, likely leads, and crystals) with the help of Dr. Pusey at iXpressGenes, Inc. Each category has its own specific characteristics that needs to be considered independently. In this paper, we focus on “crystals” only and propose a solution to select the best thresholding technique for each image.

The rest of the paper is structured as follows. Our dataset and image binarization techniques are described in Section 2. Our approach to select the best binarization technique is explained in Section 3. Experimental results are provided in Section 4. Finally, our paper is concluded with the last section.

2. Background

2.1 Dataset

We group protein crystallization images into three main categories: noncrystals, likely leads, and crystals. In this study, we focus on only images containing protein crystals and try to determine the best threshold method for images of crystal subcategories. The protein crystal images may be split into 5 main categories: “Posettes and Spherulites”, “Needles”, “2D Plates”, “Small 3D Crystals”, and “Large 3D Crystals”. Distinctive features of these categories may be identified as high intense regions, straight edges, and proper geometric shapes. Our crystal dataset set consists of 3 subcategories: 2D plates, small 3D crystals, and large 3D crystals.

2.1.1 2D Plates

The images in this category have quadrangular shapes and have 2 dimensions. In some specific cases, we may not be able to observe all the edges of a quadrangular shape because of focusing issues. 2D Plates may have small or large sizes, and they may be located as a stack of regions. The intensities of 2D Plates are lower than the intensities of 3D crystals. This means intensity change between the foreground and the background may not be as significant as for 3D crystals. Figure 1 shows a group of sample images for this category.

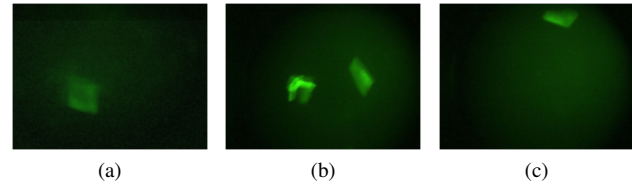


Fig. 1

2D PLATE CRYSTAL SAMPLES

2.1.2 Small 3D Crystals

The areas of small 3D crystals are smaller than those of large 3D crystals. They have higher intensities than 2D plates. This causes a significant intensity change between 3D objects and background in images. Generally, it is hard to detect all the edges of this category due to small size. Figure 2 shows some sample images of this category.

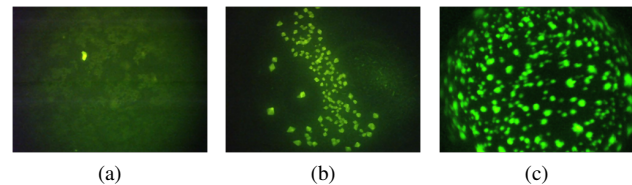


Fig. 2

SMALL 3D CRYSTAL SAMPLES

2.1.3 Large 3D Crystals

This category generally has regions with high intensity. The 3D structure of large 3D crystals can be observed in images and they have more than 4 edges. In some particular cases, it is difficult to detect all the edges because of focusing and light reflection problems. The instances of this category have larger sizes than small 3D crystals. Some sample images of this category are shown in Figure 3.

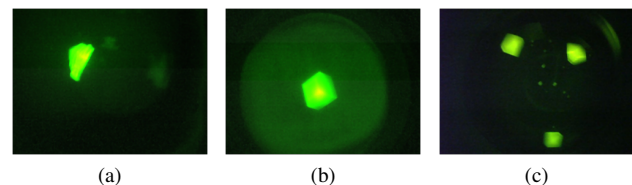


Fig. 3

LARGE 3D CRYSTAL SAMPLES

2.2 Image Binarization Methods

Image binarization is a technique for separating foreground and background regions in an image. For the pro-

tein images consisting of crystals, the crystal regions are expected to be represented as the foreground in the binary images. While a thresholding technique may perform well for an image, it may not perform as good as other thresholding techniques for another image. Thus, we consider three image binarization techniques described below.

2.2.1 Otsu Threshold

For Otsu's thresholding [5], firstly a gray level image is generated from an input color image. Then, for each possible intensity threshold, the variance of spread of pixels in the foreground and background region is calculated. The intensity (τ_0) for which the sum of foreground and background spreads is minimal is selected as the threshold. Pixels with gray level intensity higher than (τ_0) form the foreground region while the remaining pixels form the background.

2.2.2 90th Percentile Green Intensity Threshold (g90)

When green light is used as the excitation source for fluorescence based acquisition, the intensity of the green pixel component is observed to be higher than the red and blue components in the crystal regions [4]. This method utilizes this feature for image binarization. First, the threshold intensity (τ_{g90}) is computed as the 90th percentile intensity of the green component in all pixels. This means that the number of pixels in the image with the green component intensity below this intensity constitutes around 90% of the pixels. Also, a minimum gray level intensity condition ($t_{min} = 40$) is applied. All pixels with gray level intensity greater than t_{min} and having green pixel component greater than (τ_{g90}) constitute the foreground region while the rest constitute the background region [6].

2.2.3 Maximum Green Intensity Threshold (g100)

This technique is similar to the 90th percentile green intensity threshold described earlier. In this method, the maximum intensity of green component (τ_{g100}) is used as the threshold intensity for green component. All pixels with gray level intensity greater than t_{min} and having green pixel component greater than (τ_{g100}) constitute the foreground region. The foreground (object) region in the binary image from this method is usually smaller than the foreground region from the other two techniques [6].

3. Method

In this section, first we describe the generalized form our DT-Binarize that can be used in any image binarization problem. Then we briefly define the methods used at intermediate stages of our algorithm. Finally, we provide application of this method to the protein image binarization problem.

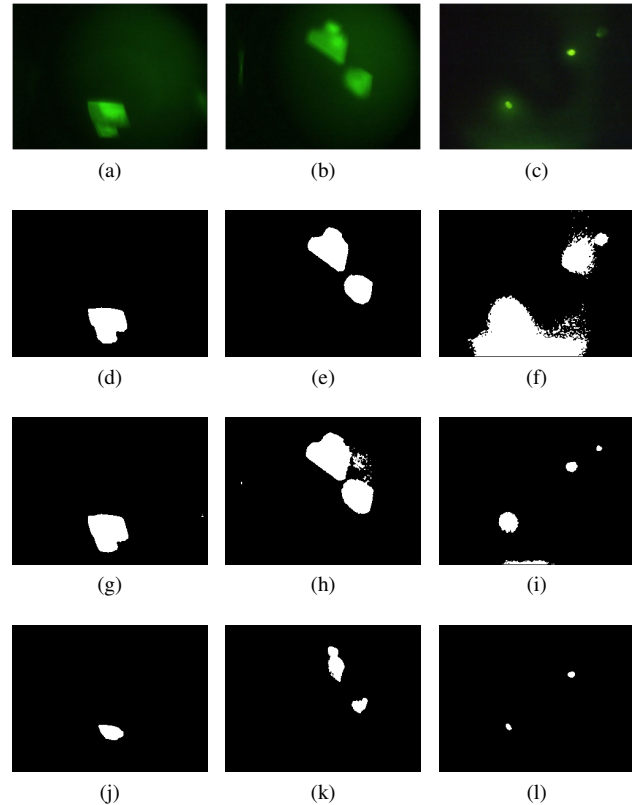


Fig. 4

EXAMPLE IMAGES THAT WORKS GOOD FOR ALL THRESHOLDING TECHNIQUES, (A), (B), (C) ORIGINAL IMAGES, (D), (E), (F) OTSU RESULTS, (G), (H), (I) G90 RESULTS, (J), (K), (L) G100 RESULTS, NOTE THAT (E),(G), AND (L) ARE THE BEST BINARY IMAGES

3.1 DT-Binarize: Selection of Best Binarization Method Using Decision Tree

Image binarization is a challenging problem. It is not practical to determine the optimal threshold value for all cases since there are some weaknesses and strengths of the all image binarization methods [7]. Based on this fact, in this research, we target an algorithm that selects the best binarization method rather than a single threshold value. Our goal is to exploit the powerful features of different binarization methods and use them whenever they perform well. For this reason, we propose using a supervised classification method using decision trees to determine the best binarization method for any image dataset based on some statistical features such as standard deviation, mean, max intensity, etc.

We first build a train set for thresholding techniques. In the training set, the best thresholding technique for each image is used as the class label. Then in the training stage, we build the decision tree based on the statistical features of the images in the training dataset. Once we have the decision

tree, we are able to determine the best binarization method for any test image by using the same statistical features. Following steps provide a brief summary of our algorithm.

- 1) Label training images with best binarization methods
- 2) Extract statistical features of the training images
- 3) Build the decision tree based on the statistical features
- 4) Predict the best binarization method for a test image using the decision tree

3.2 Stages of the Algorithm

3.2.1 Median Filter

Median filter is one of the well-known order-statistic filters due to its good performance for some specific noise types such as “Gaussian”, “random”, and “salt and pepper” noises. In median filter, the center pixel of a $M \times M$ neighborhood is replaced by the median value of the corresponding window. Note that noise pixels are considered to be very different from the median. Using this idea median filter can remove this type of noise problems [6]. We use this filter to remove the noise pixels on the protein crystal images before binarization operation.

3.2.2 Contrast Stretching

Contrast stretching is a normalization method that enhances the informative features of the image by expanding the histogram of the intensities. It maps the pixel values into a new range in a linear fashion [6]. We can apply contrast stretching to the images by using the Eq 1,

$$I_{out} = (I_{in} - P_{in}) \left(\frac{P'_{max} - P'_{min}}{P_{max} - P_{min}} \right) + P'_{min} \quad (1)$$

where I_{in} and I_{out} are the input and output images, P_{min} and P_{max} are the minimum and the maximum intensity value of the input images, and P'_{min} and P'_{max} are the minimum and the maximum intensity values of the output image, respectively. We include contrast stretching in our research, because our dataset contains some low contrast images, which causes incorrect threshold results on our dataset. Figure 5 shows a problematic image and contrast stretching result. Note that informative features of the result image are magnified without losing the structure of the crystal.

3.2.3 Decision Tree

Decision tree [8] is a rule based classifiers in the literature that employs a tree structure for data classification. It is a supervised classification technique that comprises of training and testing stages. In the training stage the tree is generated based on the entropy of the data features. In the testing stage, each test sample is classified using the tree built in the training stage. Decision tree is a classifier that requires relatively less time to create training model. Also, testing is quite fast after building the tree.

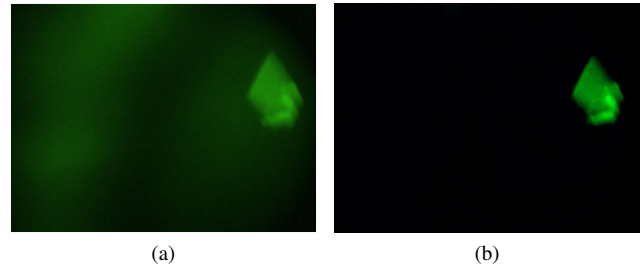


Fig. 5

CONTRAST STRETCHING EXAMPLE (A) ORIGINAL IMAGE AND (B) IMAGE AFTER APPLYING CONTRAST STRETCHING

3.3 Application to Our Problem

Protein image binarization problem is a convenient application area of our algorithm. For this specific case, we use our training images to build the decision tree based on only standard deviation of the pixel intensities. 75% of the data is selected as the training set and remaining is used for the testing. Figure 6 shows the result tree of the training stage. In Figure 6, “g90” is selected as the best binarization method if standard deviation of the test sample is less than 12.86. However, if the standard deviation is between 12.86 and 24.99, the best binarization method is selected as “Contrast Stretching + g90”. Similarly, other binarization methods may be selected depending on the standard deviation of the test image.

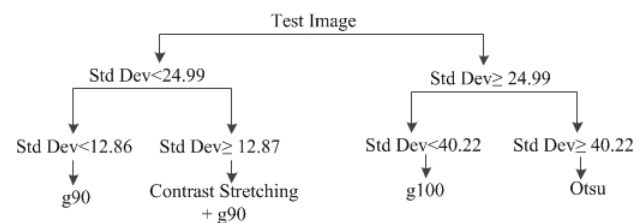


Fig. 6

DECISION TREE FOR SELECTING THE BEST THRESHOLD METHOD

We have employed this tree to our test dataset. For a test sample, we take the standard deviation and find the corresponding leaf node. The method at the leaf node is selected to binarize that test image. Following section provides some numerical and visual results of this technique with several examples.

4. Experiments & Results

This section provides objective evaluation of each binarization technique using the ground truth (reference) image dataset that is manually generated by our research group. The correctness of a binary image is calculated using several

well-known performance measures. Our DT-Binarize technique is also compared with the given methods.

4.1 Protein Crystal Dataset

Our dataset consists of totally 114 protein crystal images that consist of 3 subcategories: 2D plates (40%), small 3D crystals (10%) and large 3D crystals (50%). The size of each image is 320×240 , and all images have been captured by a special imaging system under green light. While some of the images have distinctive features such as high intensity or clear border, some of them may have unclear shapes that are difficult to differentiate crystals from the background.

4.2 Correctness Measurement

Since a simple visual comparison of the binary images of each method would not provide an objective and dependable results, we decided to generate reference (ground-truth) binary images of each sample in our dataset. So we manually extract the protein instances using an image editing software [9] that has the capability of auto selection of the objects on the image. Also we were able to adjust fine level changes on the object areas. This helps us generate ground-truth images.

Once we have the reference images, our comparison can be achieved objectively. Basically we take an output binary image and the corresponding reference binary image then measure the similarity between two images by “weighted sum” of the images. Suppose the pixels of protein instances are represented by “1” and the background area is represented by “0” in the images. In order to find out the correspondence between the images, we can use the following equation,

$$I_S = 2 \times I_R + I_O \quad (2)$$

where I_S , I_R and I_O are the sum image, reference binary image, and the output binary image, respectively. Figure 7 shows an example sum image that includes 4 regions. Note that if the pixel p_{ij} of the sum image is “3”, it is a *hit*, which is also called as a True Positive (TP). If the pixel is “2”, it is a *miss*, which is called as a False Negative (FN). Similarly, if the pixel is “1”, it is a *false alarm*, which is called as a False Positive (FP). Finally if the pixel is “0”, it is a *correct reject*, which is called as a True Negative (TN). We can use these 4 values (TP, TN, FN, TN) to measure the correctness of the output binary image.

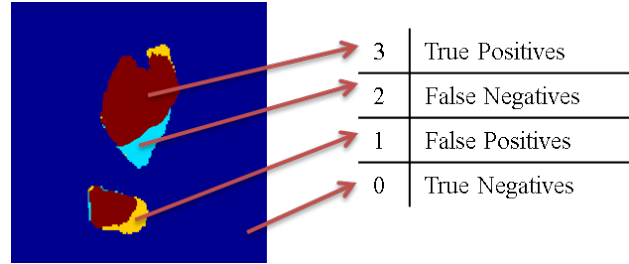


Fig. 7

EXAMPLE SUM IMAGE

In the literature there are several measures that may provide correctness information from different perspectives. It is important to use a proper accuracy measure that is more relevant to the characteristics of our study. For example, the classical accuracy measure may not be proper technique for our study. Because in a typical protein binary image, there are usually very few number of foreground pixels compared to the background pixels. This means that the TN pixels can easily suppress the accuracy even if there are no TP pixels. To avoid bias towards a specific measurement method, we use and compare 4 well-known measures: Accuracy, F-Score (F-measure), Matthews correlation coefficient (MCC), and Jaccard (Jacc) similarity. These can provide more reliable measures for a variety of confusion matrices [10]. Following equations show the formula of each measurement.

$$M_{acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$M_{F1} = \frac{2 \times TP}{(2 \times TP) + FP + FN} \quad (4)$$

$$M_{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

$$M_{Jacc} = \frac{TP}{TP + FP + FN} \quad (6)$$

4.3 Results

In the experimentation stage we generate 4 binary images using 3 binarization techniques ($g90$, $g100$, Otsu) and our algorithm (See Figure 9). Correctness of each binary image is measured based on reference binary images (ground truth). 4 different correctness measures are employed at this stage in order to evaluate the results objectively. This process is done for all test images in the dataset. Table 1 shows the average results of each measure. According to the results, our method outperforms all other methods by 10% on the average.

A visual representation of the results is given in Figure 8.

Our technique can generate the best binary image in almost all cases. Figure 9 shows a sample test case in which

Table 1
COMPARISON OF THE TECHNIQUES BY DIFFERENT MEASURES

	G100	G90	Otsu	DT-Binarize
Acc	0.9787	0.9569	0.8911	0.9844
F1	0.6935	0.6230	0.6212	0.8106
MCC	0.7184	0.6632	0.6516	0.8236
Jaccard	0.5907	0.4960	0.5396	0.7103

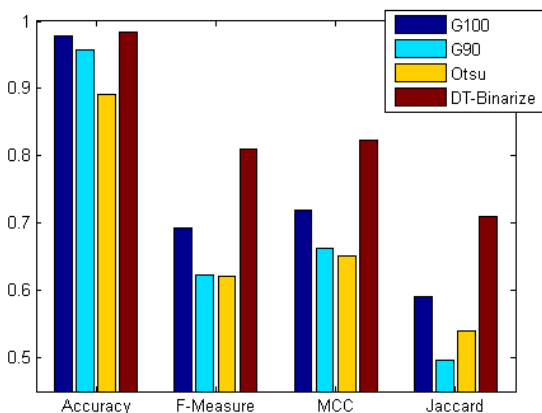


Fig. 8
COMPARISON RESULTS

our technique can successfully generate the best result. Our DT-Binarize method can adapt different lighting and focusing conditions.

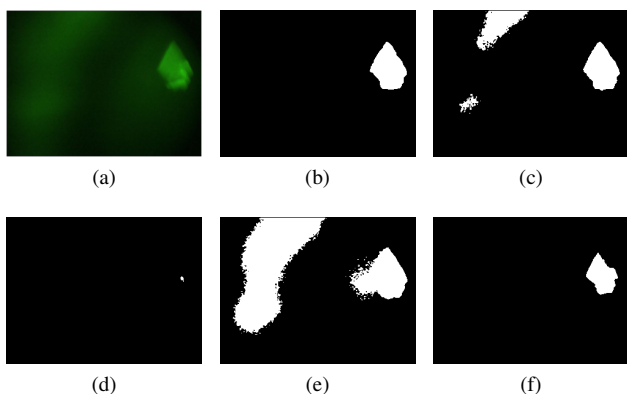


Fig. 9
A SAMPLE TEST CASE (A) ORIGINAL IMAGE, (B) GROUND TRUTH IMAGE, (C) g_{90} THRESHOLD, (D) g_{100} THRESHOLD, (E) OTSU THRESHOLD, (F) DT-BINARIZE

However, there are also a few cases that our technique could not provide accurate binary image of the protein crystal. Figure 10 shows a sample image for that case.

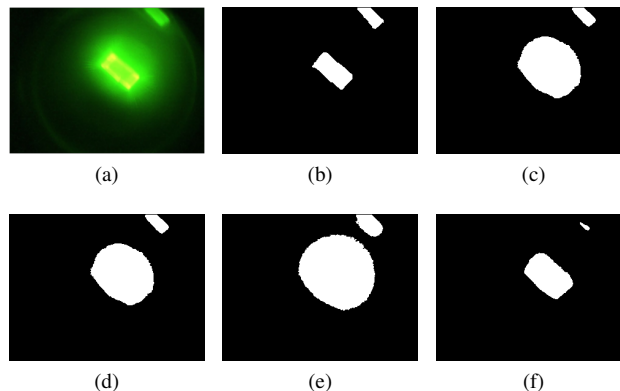


Fig. 10
EXAMPLE OF A BAD BINARIZATION RESULT (A) ORIGINAL IMAGE, (B) GROUND TRUTH IMAGE, (C) g_{90} THRESHOLD, (D) g_{100} THRESHOLD, (E) OTSU THRESHOLD, (F) DT-BINARIZE

Please note that none of the other 3 thresholding techniques can generate satisfactory binary images for these problematic samples. In other words, if none of the provided thresholding techniques provides a correct result, our method will not provide a good result either. So the performance of our method depends on the performance of the input thresholding methods. We may measure the performance of our method with respect to whether the best out of these three methods is chosen or not. As in the preparation of the training set, the best method is chosen for each image by an expert. In this case, our method is considered to perform well when it chooses the same thresholding technique chosen by the expert for each image. Figure 11 shows the comparison of the correctness of our technique with respect to expert labeling. The closeness to the limit indicates the success of our approach in this problem.

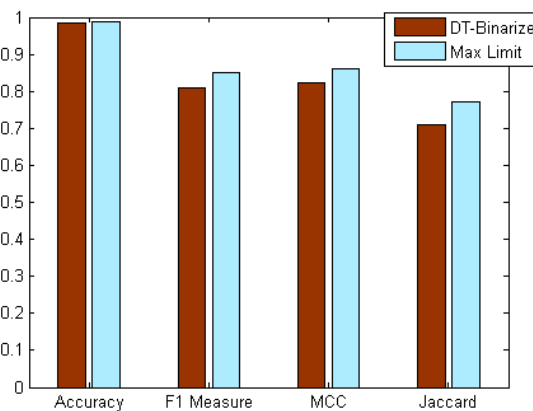


Fig. 11
COMPARISON WITH THEORETICAL MAX LIMIT

5. Conclusion

This paper presents a new technique for image binarization problem using a group of different thresholding methods. Our approach is a supervised method with training and testing stages. In the training stage, a decision tree is built using the standard deviation of the protein images. Leaf nodes of the tree represent different thresholding techniques that provide the best binarization method for a specific group of images. In the testing stage, using the decision tree, we select the best thresholding technique for the test sample and then generate the binary image using that technique.

We evaluate the performance of our approach with 4 different accuracy measures. For all cases, our method outperformed other single thresholding methods. According to the results our technique improves the binarization accuracy by 10% on the average and provides high accuracy by reaching the 95% of the expert choices.

6. Acknowledgement

This research was supported by National Institutes of Health (GM090453) grant.

References

- [1] X. Zhu, S. Sun, and M. Bern, "Classification of protein crystallization imagery," in *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE*, vol. 1, Sept 2004, pp. 1628–1631.
- [2] B. Rupp and J. Wang, "Predictive models for protein crystallization," *Methods*, vol. 34, no. 3, pp. 390 – 407, 2004, macromolecular Crystallization. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1046202304001203>
- [3] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, 2004. [Online]. Available: <http://dx.doi.org/10.1117/1.1631315>
- [4] M. Sigdel, M. L. Pusey, and R. S. Aygun, "Real-time protein crystallization image acquisition and classification system," *Crystal Growth and Design*, vol. 13, no. 7, pp. 2728–2736, 2013. [Online]. Available: <http://pubs.acs.org/doi/abs/10.1021/cg3016029>
- [5] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.
- [6] R. Gonzalez and R. Woods, *Digital Image Processing*. Pearson/Prentice Hall, 2008. [Online]. Available: <http://books.google.com/books?id=8uGOnjRGEzoC>
- [7] S. Roy, S. Saha, A. Dey, S. Shaikh, and N. Chaki, "Performance evaluation of multiple image binarization algorithms using multiple metrics on standard image databases," vol. 249, pp. 349–360, 2014. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-03095-1_38
- [8] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.
- [9] C. Corporation. (2014) Corel draw. [Online]. Available: <http://www.corel.com/corel/category.jsp?rootCat=cat20146&cat=cat3430091>
- [10] M. Sigdel and R. S. Aygün, "Pacc - a discriminative and accuracy correlated measure for assessment of classification results," in *Proceedings of the 9th International Conference on Machine Learning and Data Mining in Pattern Recognition*, ser. MLDM'13. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 281–295. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-39712-7_22

Reliability Measure of 2D-PAGE Spot Matching using Multiple Graphs

Dae-Seong Jeoune¹, Chan-Myeong Han², Yun-Kyoo Ryoo³, Sung-Woo Han⁴, Hwi-Won Kim⁵,
Wookhyun Kim⁶, and Young-Woo Yoon⁶

¹Department of Media Design, Daegu Future College, Kyungsan, Kyungbuk, Republic of Korea

²M+VISION Co. Ltd., Daegu Metropolitan City, Republic of Korea

³Department of Medical Computer Science, Daegu Health College, Daegu, Republic of Korea

⁴Department of Computer Engineering, Daegu Science University, Daegu, Republic of Korea

⁵Education Center for IT, Kyeongbuk College, Yeungju, Kyungbuk, Republic of Korea

⁶Department of Computer Engineering, Yeungnam University, Kyungsan, Kyungbuk, Republic of Korea

Abstract - This paper proposes the reliability measure scheme of 2D-PAGE spot matching based on the grassfire spot matching algorithm by introducing multiple graphs, whereas the previous research on spot matching took advantage of a single graph. Since the result of spot matching contains errors such as mismatching and false-positive matching, all of the matched spot pairs should be examined one by one to be confirmed. In the previous research on spot matching, either overall or randomly selected spot pairs from the spot matching result have to be inspected for final confirmation because it does not present data concerning matching reliability. In this paper, however, the matching reliability by accumulating individual matching result using multiple graphs is presented. As a consequence, the effort for verifying matching result in manual can be significantly reduced, and moreover which positively influences on the overall matching accuracy.

Keywords: Matching Reliability, Spot Matching, 2D-PAGE, Grassfire Algorithm, k-NNG, Multiple Graphs

1 Introduction

In proteomics, the two-dimensional polyacrylamid gel electrophoresis(2D-PAGE) is widely used to separate and identify proteins. The principle of 2D-PAGE is to separate proteins in a sample on two dimensional gel plane by taking advantage of two distinct characterizing features such as iso-electric point and the mass of each protein. A resulting gel image after performing 2D-PAGE contains a number of spots representing separated proteins with different size, shape and location[1].

In the research of proteomics, it is very important to study the expression of individual protein in a specific tissue. Even in the experiment with samples of the same tissue, proteins are to be expressed differently according to environments. In order to trace this kind of different protein expressions, the reference gel of normal state is compared to the target gel of test sample. It is very hard or almost

impossible to carry out this process in manual because hundreds or thousands of proteins are generally contained in a gel. Therefore, it is necessary to automate the analytical process of 2D-PAGE[2].

The experiment process of 2D-PAGE is straightforward, but involves experimental errors. The two experiments in the same laboratory even with the same sample, equipment and environment show considerable differences, especially in the displacement between the locations of the corresponding spots. This kind of low repeatability is main cause of making the automated analytical process of 2D-PAGE more difficult. So, many algorithms to resolve the problem and reflect the characteristics have been proposed. The automation process is commonly divided into two phases: spot detection and spot matching. The former is to separate for identifying spots from the background in a 2D-PAGE gel image, whereas the latter is to match the same or similar spot pairs in their locations between reference and target gels.

In this paper, a novel method on how to measure the reliability of the spot matching results is proposed. The first step is the basic spot matching procedure that is repeated with the grassfire algorithm[3] using multiple k-NNGs. And then, each result is accumulated to evaluate reliability of every matched spot pairs. With this scheme, researchers can verify their experiment results with reliability based on probabilistic way. At the same time, they can make decision on the range of spots that are manually verified. Therefore, the reduction of cost and time required for result verification in 2D-PAGE can be achieved.

2 Spot matching in 2D-PAGE

2.1 Definition

Let the spot sets P and Q be reference and target gels, respectively. Every element of P and Q has its own two-dimensional coordinate representing the central point of spot as $p_i=(x_i, y_i)$ and $q_j=(x_j, y_j)$. Here, the spot matching is to find out the set of matched spot pairs and simply defined as M

between a reference gel P and a target gel Q that satisfy the given conditions[4], as shown in the equation (1).

$$\begin{aligned} P &= \{p_1, p_2, \dots, p_n\}, \text{ where } p_i = (x_i, y_i), 1 \leq i \leq n \\ Q &= \{q_1, q_2, \dots, q_m\}, \text{ where } q_j = (x_j, y_j), 1 \leq j \leq m \\ M &= \{(p_{i1}, q_{j1}), (p_{i2}, q_{j2}), \dots, (p_{il}, q_{jl})\}, \\ &\text{ where } p_{il} \in P, q_{jl} \in Q \text{ and } l \leq \min(m, n) \end{aligned} \quad (1)$$

2.2 Spot matching algorithm

As the 2D-PAGE gel images have characteristics of not only showing low repeatability of experiment process but also both including global and local distortions, various spot matching methods have been proposed. One of them is known as image matching approach that has been proposed in order to resolve spot matching problems using reference and target gels involving nonlinear distortions[5-6]. And another is point pattern matching that utilizes geometric property of a graph transformed with respect to central positions of spots from gel image. So far, the spot matching algorithm using the geometric property can be classified into several categories. There include spot matching methods using the landmark spots, graph theory, iterative closest point, similarity the neighboring spots, etc.

2.3 Grassfire algorithm

In this paper, *the grassfire algorithm*[3] that is based on the topological pattern of neighboring spots in a graph is adopted to perform basic spot matching with 2D-PAGE images. The grassfire algorithm is characterized by three important factors: selection of graph type, determining the initial point called the seed spot, and the direction of spot matching. The algorithm shows different result according to the type of a graph. Among the commonly used graphs in spot matching such as the Delaunay triangulation graph, the Gabriel graph, the relative neighbor graph and the k-NNG (nearest neighbor graph), the last one is used in the grassfire algorithm.

Spot matching in the grassfire algorithm starts at a single matched spot pair that is confirmed, which is called the *seed spot pair*. It should be clearly determined as a positive matching spot pair, otherwise unreliable matching result can be produced due to false candidates. In general, the seed spot pair is determined in manual or by an algorithm using the landmark spots. As a next step, a single spot pair with most similarly matched is selected among the neighboring spots of the seed spot to perform spot matching.

The grassfire spot matching scheme shows good performance in accuracy and speed because it takes advantage of matching information of the previous matching stage at the next stage. And spot matching between spot pairs is spread toward the direction of showing higher matching accuracy because the next matching candidate is determined as a spot

pair of the highest topological similarity among neighboring spots. Also, the position of seed spot that is closely related to the direction influences on matching order because matching starts at that point. Nevertheless, the grassfire algorithm shows the same matching result regardless of its position because a spot pair with best matching score among neighboring spots in the previous stage is selected as the next matching candidate.

3 Reliability measure

3.1 Verification of spot matching result

In spot matching algorithm, accuracy is the most important. Therefore, it is necessary to verify spot matching result by the automated method because the accuracy of 100% is not guaranteed. The only way for verification is absolutely processed by human in manual, but how to verify the spot matching result is not a trivial thing. When verifying spot matching result at random, there is possibility not to involve the false- matched spot pairs in the verification candidates. In case overall matched spot pairs are examined, it requires so much time and cost.

If the matching reliability for each matched spot pair after spot matching can be presented, one can easily determine the range of spot pairs that should be verified from the overall matched spot pairs. And then, all the spot pairs within a certain range are verified for confirmation accurately. Let us, i.e., suppose that all the spot pairs whose reliability is less than or equal to 30% is verified in manual. When every spot pair is to be correct, each of the remaining spot pairs with reliability of higher than 30% is likely to be confirmed and regarded as correct matching pairs. To the contrary, when there exists spot pair(s) that turns out to be incorrect, the reliability for verification should be raised, for instance, up to 50%. In this manner, manual verification can be done with minimum cost with no verification for the overall spot pairs in the spot matching result by algorithm.

3.2 Reliability measure using multiple graphs

Once the grassfire algorithm is executed, then a set of matched spot pairs is obtained as a spot matching result. The matched spot pair is a tuple that consists of number of the spot in the reference and the target gels, forming (p_i, q_j) . The performance of the grassfire spot matching algorithm is influenced only by the set of spots in reference and target gels and the type of graph. So, the result of the grassfire scheme can be expressed as $grassfire(P, Q, G)$.

In this paper, multiple k-NNGs are applied in turn to the same sets of the reference gel P and the target gel Q for each iteration of the grassfire spot matching algorithm. The result for each k-NNG is produced and accumulated in the accumulated frequency matrix (AFM). The AFM is two-dimensional array and its X- and Y-axes consist of spot number in the reference and target gels, respectively. When the spot pair (p_i, q_j) is included in the result, then the array $AFM(p_i, q_j)$ is increased by 1.

When executing the grassfire algorithm, the k-NNG is used by varying the degree $k=5$ to 8 with step size as 1 in turn because the number of neighboring spots in the same kind of graph is sequentially increased, which holds the similar property between graphs and, at the same time, the difference between them. The similarity helps the algorithm to determine the distinct matched spot pair by allowing higher frequency. And the difference contributes accurate matching rate enhancement for the uncertain spot pair with no topology conservation by allowing attempt to matching with several different spots. When four kinds of graphs are used in the spot matching and the frequency stored in the AFM for a specific spot pairs is four, the reliability for that matched spot pair becomes 100% . In other words, it means that the spot pair is positively matched because the matching for it is included in the results of every spot matching using four different graphs.

4 Experiment and result

4.1 Data set for experiment

A pair of reference and target gels whose matching result is already known is needed to test and verify the proposed reliability measure scheme. However, it is very time-/cost-consuming process to prepare samples after confirming every matching spot pairs in manual because hundreds or thousands of spots are contained in a single gel. Therefore, a pair of gels is generated using simulation method for convenience in this paper.

First of all, 25 spots in a gel with the size of 64×64 pixels are randomly generated and it is regarded as a reference gel. In random process of spot generation, the minimum distance between spots should be limited lest they should be overlapped or placed too close. After the spots of a reference gel are generated, they are transformed to produce spots for a corresponding target gel. This paper uses random number of the normal distribution to produce the target gel from the reference gel. The displacements for every spots in the

reference gel is calculated using the normal distribution to the X- and Y-axis, respectively. And then, they are added to the coordinates of the corresponding spots in the reference gel, resulting in the target gel. Here, it is assumed to have no outlier spots for simplification.

As a next step, the same number is assigned to the matched spots in the reference and the target gels. It provides easy way to identify the matched spot pairs of the matching result by spot matching algorithm whether each pair belongs to false-positive or true-positive with pre-assigned spot numbers. The Figure 1 shows an example of the generated gel data for experiment to verify the proposed scheme.

4.2 Experiment and result

For brevity and clarity of experiment, a pair of gels with 25 spots is used in the experiment, even though hundreds or thousands of spots exists in a gel in real. The procedure of experiment is straightforward. First, Spot matching using the grassfire algorithm using a single graph of k-NNG from $k=5$ to 8 is performed one by one. And then, the matching result using the grassfire algorithm with each graph is accumulated in the AFM, as described in the previous chapter. The spot matching performance for each graph is summarized in the Table 1. In the table, the spot detection rate is ratio of the number of the matched spots over that of overall spots in a gel. And, the accuracy is ratio of the number of the true-positive matched spot pairs over that of overall matched spot pairs. After all, it is shown that the false-positive spot pairs are included in the matching results for the k-NNGs with degree $k=5$, $k=6$, and $k=7$.

As the next step, the reliability measure is performed using the spot matching results with multiple graphs. When the result using the grassfire algorithm with 5-NNG is produced, the value of 1 is stored to the corresponding position in the AFM for every matched spot pairs. In turn, the same procedure is performed except that the graph is changed to 6-NNG, 7-NNG, and 8-NNG in turn, and the array of the AFM

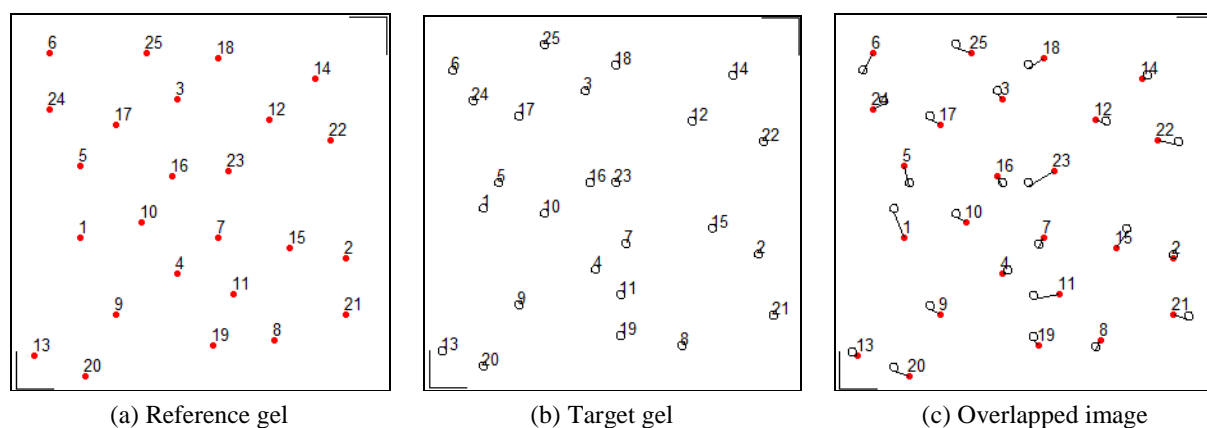


Figure 1. The generated gel image data for experiment (image size of 64×64 pixels, 25 spots, and the minimum distance of 10 pixels)

is accumulated by increasing 1 for the every matched spot pairs. Finally, the matching reliability is calculated for every matched spot pair using the information in the AFM.

The result of the reliability measure using multiple graphs is shown in the Table 2. If the matched spot pairs whose matching reliability is higher than 50% are determined to be matched, both the spot matching detection rate and the accuracy become 100%. The matched spot pairs (24-6), (4-11), (12-23) and (23-12) are false-positive matching spot pairs

resulted from the grassfire spot matching algorithm using 5-NNG, 6-NNG and/or 7-NNG. In the proposed multiple graph method, the inaccurate matched spot pairs can be excluded by calculating matching reliability for all the matched spot pairs. Actually, the spot pairs matched by chance in a certain repetition of the grassfire algorithm shown in the shaded rows of the Table 2 show very low reliability compared to the others.

Table 1. Grassfire spot matching result using multiple k-NNGs

Measures	Matched spot pairs			
	k=5	k=6	k=7	k=8
Detection rate(%)	92.0	96.0	100	100
Matching accuracy(%)	95.7	95.9	92.0	100

Note : The value k denotes the degree of k-NNG in each repetition using the grassfire spot matching algorithm.

Table 2. Reliability measure for the matched spot pairs using multiple k-NNGs

Spot pairs (i, j)	Matched spot pairs				Cumulative matching frequency	Matching reliability (%)
	k=5	k=6	k=7	k=8		
(1-1)	√	√	√	√	4	100.0
(2-2)	√	√	√	√	4	100.0
(3-3)	√	√	√	√	4	100.0
(4-4)	√		√	√	3	75.0
(5-5)	√	√	√	√	4	100.0
(6-6)		√	√	√	3	75.0
⋮	√	√	√	√	4	100.0
(11-11)			√	√	2	50.0
(12-12)	√	√		√	3	75.0
⋮	√	√	√	√	4	100.0
(23-23)	√	√		√	3	75.0
(24-24)		√	√	√	3	75.0
(25-25)	√	√	√	√	4	100
(24-6)	√				1	25.0
(4-11)		√			1	25.0
(12-23)			√		1	25.0
(23-12)			√		1	25.0

Note : A spot pair (i, j) denotes represents a matched spot pairs with a spot number in the reference gel i and that of in the target gel j.

5 Concluding remarks

In the field of proteomics, the spot matching algorithm has been devised to automate one or more 2D-PAGE phases, which reduces not only time and cost but also endeavor of researchers. However, every spot matching result from the automated method should be examined manually one by one with large amount of time and cost because it contains errors such as false-positively matched spot pair(s).

For this reason, the reliability measure using multiple graphs has proposed in this paper. The proposed scheme can provide a method of reliability measure for each of the matched pair using the grassfire spot matching algorithm with multiple graphs. And then, the experiment is performed to verify the effectiveness using the generated 2D-PAGE data images with 25 spots in a 64×64 pixels. From the experiment result, it has shown that enhancement of spot matching detection rate can be achieved by examining every reliability of the whole matched pairs to determine the appropriate threshold on reliability. And, it helps to minimize the time and cost of manual verification process that should be done in the automated process. Moreover, it is far more cost-effective when many of reference gels are to be compared with a single target gel.

The contribution of this study can give new opportunity to use spot matching algorithm more in practical use because it helps to increase manual verification efficiency with the minimum cost. The further research includes the automatic determination of reliability threshold by analyzing the frequency distribution histogram of matching reliability.

6 References

- [1] P. H. O'Farrel, "High Resolution Two-Dimensional Electrophoresis of Proteins", *Journal of Biological Chemistry*, Vol. 250, No. 10, pp. 4007-4021, May 1975.
- [2] T. Srinark and C. Kambhamettu, "An Image Analysis Suite for Spot Detection and Spot Matching in Two-Dimensional Electrophoresis Gels", *Electrophoresis*, Vol. 29, pp. 706-715, 2008.
- [3] Yun-Kyoo Ryoo, Chan-Myeong Han, Ja-Hyo Ku, Dae-Seong Jeoune, and Young-Woo Yoon, "Grassfire Spot Matching Algorithm in 2-DE", *International Journal of Bio-Science and Bio-Technology*, Vol. 5, No. 4, pp. 167-174, 2013.
- [4] Chan-Myeong Han, Dae-Seong Jeoune, Hwi-Won Kim, and Young-Woo Yoon, "A Spot Matching Algorithm using the Topology of Neighbor Spots in 2D-PAGE Images", *International Journal of Software Engineering and Its Applications*, Vol. 7, No. 5, pp. 87-97, 2013.
- [5] Jiann-Der Lee and Wei-Chun Chen, "A Novel Scheme for Registration of Two Dimensional Gel Electrophoresis Images",

Biomedical Engineering: Applications, Basis and Communications, Vol. 18, No. 4, pp. 158-166, August 2006.

- [6] G. Shi, T. Jiang, W. Zhu, B. Liu and H. Kao, "Alignment of Two-Dimensional Electrophoresis Gels", *Biochemical and Biophysical Research Communications*, Vol. 357, pp. 427-432, 2007.

Active Surveillance in Public Environments

Timothy Sweet and Mircea Nicolescu

Department of Computer Science and Engineering, University of Nevada, Reno, Nevada, USA

Abstract—*The growing interest in automated surveillance in the context of homeland security, civilian, and military applications is fueling demand for robust and efficient surveillance software. This capability is particularly desirable in public environments such as federal buildings or airports where human operators have to monitor a large number of camera feeds. Humans have been known to be unreliable and inconsistent in this multitask-oriented environment. In this paper, we present a framework for automatically detecting several types of interactions involving human agents and objects they carry. The proposed system uses a novel combination of vision-based techniques that allow real-time detection, as well as offline searching for instances of events with specific attributes in a large video. We validate the system on a set of sample video sequences exhibiting common events we expect to see in public environments.*

Keywords: surveillance; computer vision; tracking

1. Introduction

Interest in automated surveillance has shown a significant growth as the government, military, or private companies seek to deploy robust vision-based surveillance systems in a wide range of environments [1], with applications that include behavior monitoring and theft identification, both in real-time and offline. Solutions have ranged from simple intruder detection [2] to complex tracking of abnormal behavior of individuals among crowds [3]. We propose a system which detects relevant activities/events in complex scenes such as federal buildings, airports, shopping malls, and other public places. By developing a method to reliably detect specific actions, security personnel no longer need to continuously watch video monitors and can focus on less monotonous, more important security tasks. The system receives video from one or many cameras and tracks people and objects over time, and can alert a human operator when a certain activity is identified.

Prior work in autonomous video surveillance systems vary largely in implementation, application, and sensor models [4] [5]. There are few examples of autonomous video surveillance systems deployed in real settings. Currently, security personnel in a sensitive public facility such as a federal building or airport use a large number of monitors to manually scan the live video for suspicious events [1] [4]. This repetitive, monotonous, and multitask-oriented environment is not well suited for human attention [6]. Basic identification of relevant events involving human agents and objects they

carry can instead be tasked to an automated system, which can accomplish this task more quickly and sometimes more accurately [5], and free the security personnel to investigate suspicious activity when alerted by the system.

Possible applications of the proposed system fall into two general categories: live detection and offline review. Live detection corresponds to the typical job of security personnel: watching for any instance of suspicious activity as it unfolds. Offline review occurs when security or law enforcement personnel receive information about a past event, and wish to see surveillance footage from the event. For example, a witness to a crime may report that they saw a person with a black shirt and blue pants walking away with a brown bag. The task of searching in surveillance footage from a large array of cameras to find such a specific event is well suited for our system.

Our previous work [7] provided a high-level description of a modular system for analyzing human behavior in the context of security applications. This paper is focused on the specific vision-based capabilities for such a system and presents a novel combination of various methods into a practical, functional system for monitoring interactions involving humans and relevant objects. Our system is designed to be applicable to public surveillance scenes such as airports.

The rest of this paper is organized as follows. Section 2 presents previous work related to vision-based surveillance and activity recognition. Section 3 provides a technical description of the methods used in segmentation and tracking. Section 4 presents the approach used for event detection. Section 5 provides an experimental evaluation of the system. The paper is concluded in Section 6.

2. Related Work

Räty provided a survey of the state of the art in surveillance systems in [5]. This describes the “3rd generation” multisensory systems which combine information from multiple cameras to produce intelligent information. Common applications of multisensory surveillance systems include systems to detect intruders or track objects. Other relevant surveillance systems are described in [8] [9] [10] [11]. Haritaoglu et al. [12] demonstrated a system for detecting people by predicting positions of body parts. This led to the ability to track people in groups and across occlusions. Their work focuses exclusively on monochromatic, low resolution images, and relies on head detection for groups of people. This assumes a nearly horizontal view of the scene, and

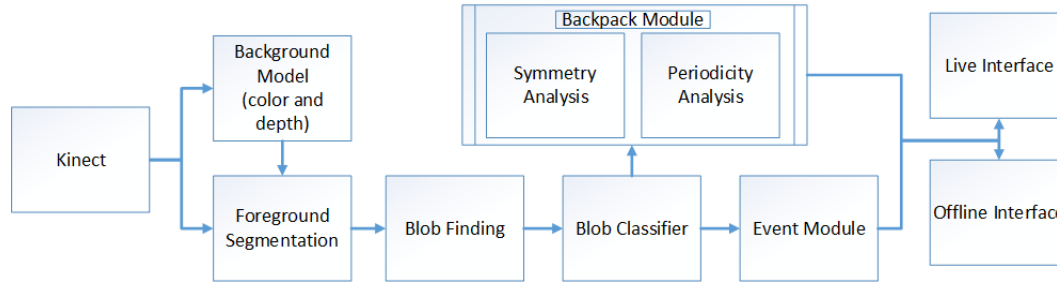


Fig. 1: The process architecture of our system

is not suited well for security cameras which are generally mounted on or near ceilings such that they are looking down on a crowd of people.

Zhao et al. [13] presented a method for detecting and tracking multiple humans outdoors by explicitly providing the camera's extrinsic parameters and time-of-day/weather information (for shadow removal). Their work shows promising results in human body detection but requires a non-trivial setup in physically calibrating the camera and environment to be observed, while also using non-vision information such as sun position and cloud coverage. This becomes prohibitively complex and time consuming with security cameras numbering in the thousands at large public scenes. When considering possible environment changes in a dynamic scene such as an airport, modeling a scene becomes impractical and a vision system designed for surveillance should be able to adapt automatically to these changes.

Related to the capabilities proposed in our system that involve interactions with objects, Cutler and Davis demonstrated in [14] a method for identifying periodic motion using low resolution video by Fourier analysis. This work was extended by [12] to detect baggage carried by a person.

3. Segmentation and Tracking

We use a Microsoft Kinect for Xbox camera to capture information about a scene. The Kinect provides a color image at VGA quality (640 x 480 pixels) as well as a depth image at VGA quality which provides information about the distance of objects from the camera. Both data streams are processed at 30 frames per second. The color and depth images are both utilized to obtain a robust image segmentation. The Kinect in particular is not critical to this application: Zhao et al. presented in [13] a method for segmentation and depth-position computation under the assumption of a static camera and known ground plane. Images are processed using the Open Source Computer Vision (OpenCV) library.

The block diagram in Fig. 1 shows the process architecture of our system. In the first stage, the system uses both color and depth information to segment images into background and foreground regions. In the second stage, the foreground pixels are grouped into blobs, and the identity of each blob is maintained across frames. In the third stage, the

blob descriptions are analyzed by two independent modules: the event module (section 4.1) and the Backpack module (section 4.2). The results of these modules are then sent to the user interface, either live or offline.

3.1 Background Modeling and Foreground Segmentation

For each frame, a color and depth image is acquired from the camera and the system performs foreground-background segmentation to determine potentially relevant foreground objects. We use a Mixture of Gaussians (MOG) model as modified by Zivkovic in [15] as the basis for background subtraction with added higher-level recovery filters.

The MOG model for background subtraction uses Gaussian intensity distributions to model the image background [15]. Zivkovic's method additionally allows detection and segmentation of shadows in the color image. Fig. 2 b-d shows the result of the color and depth background segmentations.

Some additional filtering is applied in addition to the MOG model. For every incoming color frame I with m rows and n columns the average intensity Q of the image is computed over every pixel $I(x, y)$ as:

$$Q = \frac{\sum_{y=0}^m \sum_{x=0}^n I(x, y)}{m * n} \quad (1)$$

Given the intensity of the previous frame Q_{p-1} and the current intensity Q_p an intensity scale factor is computed as:

$$F = \frac{Q_p - Q_{p-1}}{r} \quad (2)$$

where r is the history length of the MOG model. This scale factor is then broadcasted together with the color image. The effect is that sudden, global changes in lighting are smoothed across many frames and generally do not cause false positives in the foreground detection.

Occasionally, a major global lighting change, for instance a bright light bulb being turned on or off, cannot be smoothed by the intensity filter because the camera's exposure level needs time to adjust to the new lighting levels, thus there is insufficient quality data to observe from the camera (this typically manifests itself as a completely black or completely white image).

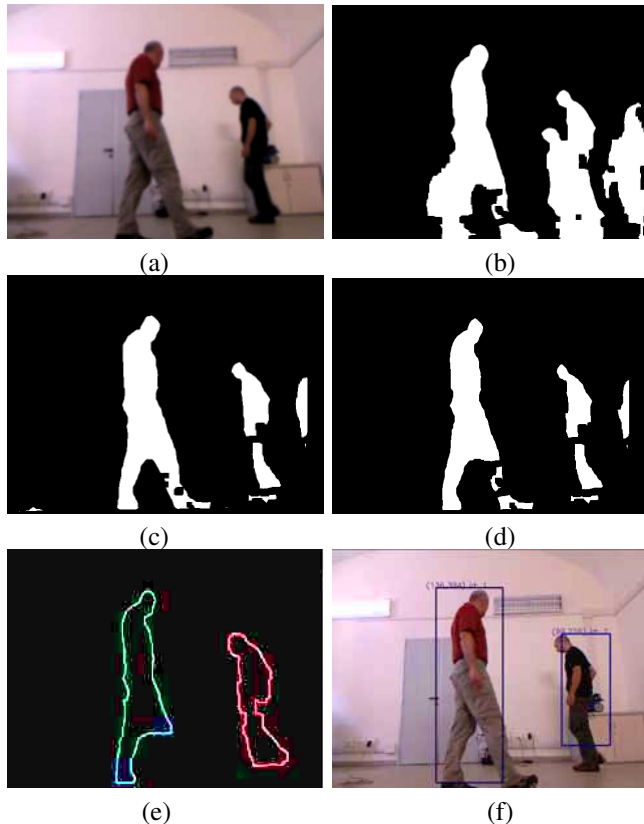


Fig. 2: The stages of blob tracking. (a) The color input from the camera; (b) the color segmentation, notice many shadows are present; (c) the depth segmentation; (d) combined segmentation; (e) detected regions are outlined in a unique color; (f) The people are marked and labeled.

During exposure adjustment, it is undesirable to continue to track objects, thus the system detects this major change and compensates by withholding new segmentation updates. Exposure adjustment is detected as a dramatic change in the size of tracked objects, i.e. the entire frame becomes foreground. The rest of the system is designed to interpret a lack of segmentation updates as a temporary failure and continue to make intelligent predictions about the movements of previously tracked objects.

In a foreground-background segmentation, the color and depth models behave quite differently. Color models tend to produce good segmentations for color-distinctive objects which are generally uniform in color. A color segmentation tends to produce false negatives when a relevant object has a similar color to the background and false positives on an object's shadow. Depth models tend to produce good segmentations for objects which are physically separated from other objects and are non-reflective. A depth model tends to produce false negatives at the points where an object touches the floor.

Thus the color and depth models have orthogonal failure modes and are combined to improve the segmentation. We perform a logical AND between the color and depth-based segmentations to produce a final segmentation.

3.2 Blob Finding

A blob B_p is a group of pixels in a locally dense area of an image which are considered to represent a single object. Pixels are initially classified into blobs using the border following approach from [16]. Partial and insignificant blobs are then merged and deleted using higher-level recovery filters described in Section 3.3.

The segmentation produced in Section 3.1 often fails in regions of the image where there is a sharp change in depth causing a shadow. One common example of this is the region where a person's shirt sleeve meets his or her arm. There will often be three distinct blobs in the torso region: the torso itself and the two arms, with the arms separated from the torso by a thin line of background pixels at a sleeve.

To merge blobs which might be part of the same object a simple region-growth algorithm is implemented. Given a set of blobs in the image, each blob is dilated by one pixel and the border following approach from [16] is repeated. If two blobs are now connected they are merged into a single, larger blob. This process is repeated until a single blob remains for a maximum of ten iterations.

We define the minimum blob size S_{min} as the minimum apparent area of a human at the far range of the Kinect's depth stream (approximately six meters), experimentally found to be 400 pixels. If a blob B_p does not satisfy:

$$area(B_p) > S_{min} \quad (3)$$

the blob is considered to be insignificant and is discarded.

3.3 Blob Classification and Tagging

Each blob in the list of significant blobs is then mapped to an existing blob or added as a new blob. At a high level, the system projects the position of each blob from the previous frames into the current frame and searches for a blob which is sufficiently similar to the known blob. The similarity comparison is performed by comparing the image histograms using the Chi-Squared method from [17]. At frame p , for blobs that have been tracked longer than 120 frames we also require the similarity between a previously observed blob B'_p and a candidate matching blob B_p to be within one standard deviation of the mean prior similarity. This extra element increases robustness to blobs occluded by similarly colored blobs. We define $x(B_p)$ as the bounding box of B_p (with no filtering). A Kalman Filter [18] is used to filter the four components of $x(B_p)$:

$$x(B_p) = \{i, j, width, height\} \quad (8)$$

where i, j is the location of the center of $x(B_p)$. At frame p the *a priori* estimate of $x(B_i)$ is updated with the current observation to compute the filtered bounding box of B_p called $\bar{x}B_p$.

This filtering affords the system greater fault tolerance by allowing tracking to appear to continue even if a blob is

temporarily not detected by the blob tracker. A temporary tracking failure might occur due to events such as a global lighting change (either due to ambient light changing or camera exposure adjustment), or occlusion by another person.

To update the previously observed blobs B , the system attempts to find a mapping between the list of currently observed blobs B_p and the known blobs B'_p using our Blob Candidate Matching algorithm in Fig. 3.

Every blob in B' is next updated given the mapping between B and B' . A blob is updated with our Blob Update algorithm in Fig. 4.

4. Event Detection

We define an event as any of six detectable situations, which include people and their interactions with each other and baggage. The Event Module detects and generates information for five of these events where the people and baggage are visually separated, while the Backpack Module does the same for baggage which is attached to a person (such as a backpack).

4.1 Event Module

For each frame, once the system has a list of currently visible blobs B' (more precisely: blobs which were matched successfully in the past 30 frames) a set of events can be identified. The blobs' histograms, filtered positions in the image plane, and average depth position are used to identify events. Five classes of events are detected, as described below. In each case, the system marks a blob or set of blobs involved in each event type in the user interface when the event is detected.

- *Person (SP)*. A user can describe a person of interest as a bimodal color distribution representing the lower and upper halves of the body, typically corresponding to shirt color and pants color. The user provides these colors as samples using either a standard HSV color picker or a previously acquired sample.
- *Two People Meeting (TP)*. A user can describe two people of interest in a similar fashion to the single person detection. A meeting between two blobs representing these people is then characterized as:
 - the distance between the blobs' geometric centers in 3D is no greater than the sum of their apparent widths in the image plane.
 - the blobs remain in this configuration for longer than one second.
 The first criteria distinguishes between people meeting versus simply being in the same frame by coincidence and provides simple scaling in apparent size due to a blob's distance from the camera, and the second criteria avoids detecting transient passing as a meeting.
- *Bag Unattended (BA)*. Baggage is detected similarly to humans. A bag is characterized as:

- For each blob B_p in B and B'_p in B' :
 - Where D_{min} is the apparent depth of a body, if:

$$center(x(B_p)) \in \bar{x}(B'_p) \text{ or } center(\bar{x}(B'_p)) \in x(B_p) \quad (4)$$
 and

$$|depth(x(B_p)) - depth(\bar{x}(B'_p))| < D_{min} \quad (5)$$
 Store B_p as a candidate match for B'_p . This ensures the two blobs are close to each other in 3D.
- For each B_p which is a candidate match for B'_p :
 - Define $U(B)$ and $L(B)$ as the histograms of the upper, lower halves of A , B , respectively, and $C(a, b)$ as the χ^2 similarity between histograms a and b .
 - Compute the similarity between B_p and B'_p called S_{B_p, B'_p} as:

$$S_{B_p, B'_p} = C(U(B_p), U(B'_p)) + C(L(B_p), L(B'_p)) \quad (6)$$
 and store it with B_p .
- For every B'_p :
 - Select the B_p with the highest S_{B_p, B'_p} value:
 - If B'_p has been tracked for less than 120 frames (typically four seconds for a 30 frames per second camera):
 - * Accept B_p as a match and update the blob (procedure described in Fig. 4).
 - Otherwise:
 - * Compute $mean(S_{B_p, B'_p})$ and $std_dev(S_{B_p, B'_p})$ from the last 120 frames, and accept the current blob as a match if:

$$S_{B_p, B'_p} > mean(S_{B_p, B'_p}) - std_dev(S_{B_p, B'_p}) \quad (7)$$
 - If B'_p is not matched with any blob in B but has been updated in the last 30 frames: add an artificial $S_{B_p, B'_p} = 0$ to its list of prior similarities. This prevents the known blob's prior similarity from becoming arbitrarily precise and allows for tracking lapses of up to one second.

Fig. 3: The Blob Candidate Matching algorithm.

For every B'_p and its match B_p :

- Store S_{B_p, B'_p} from the Blob Update algorithm.
- Update its Kalman Filter $\bar{x}(B'_p)$ with the new blob's bounding box.
- Update its $U(B'_p)$ and $L(B'_p)$ as the average of the previously observed histograms.
- Update its depth as the average of the previously observed depths.

Fig. 4: The Blob Update algorithm.

- being one of a pair of blobs whose geometric centers in 3D are no further apart than the sum of their apparent widths in the image plane.
- being less than half the height of its companion blob.

In each pair, the taller blob is flagged as the owner blob and the shorter blob is flagged as the bag blob. Following a detection of a person and a bag, an unattended bag is characterized as: an owner blob becomes separated from its bag blob according to the 3D distance metric.

- *Two people exchanging a bag (BE)*. Following the detection of an owner/bag pair and a second non-owner person blob, an exchange is characterized as:
 - an owner blob becomes separated from an owned blob while a different person blob is within owning distance of the bag blob.
- *A person stealing a bag (BS)*. Following the detection of an owner/bag pair, a bag theft is characterized as:
 - a bag is abandoned by an owner. At this instant its bounding box in the image plane is recorded.
 - the bag's geometric center leaves its recorded bounding box before the owner is reassociated with the bag.

As a user interface element, the closest person blob to a stolen bag will be reported as the thief, however the thief blob is not used for detection of the event.

4.2 Backpack Module

We implemented the periodicity analysis method from [14] as modified by [12] to detect baggage which is directly on a person's body. This method utilizes the observation that the human silhouette is mostly symmetric when walking, and its non-symmetric regions are exhibiting a periodic motion. Thus non-symmetric, non-periodic regions of the silhouette are known to be not part of the body, typically an object such as a backpack or briefcase. This class of objects is characterized as:

- **Backpack (PK)**. A backpack is a portion of a blob which is asymmetric and does not exhibit periodic motion.

The detection is implemented with our Backpack Detection algorithm in Fig. 6.

4.3 Live and Offline Interface

The six events which can be detected in live mode (five from the Event Module and one from the Backpack Module) also have a comparable offline mode. This mode is designed to be run on pre-recorded sequences in order to answer attribute-based queries such as "Did person with description X ever exchange a bag with person with description Y?" The detection algorithms are performed offline and the result is a list of timestamps of when described event occurred. Once the detection is performed once, the event timestamps are saved and can be reused for future analysis. The offline interface is shown in Fig. 7.



Fig. 5: (a) A person is tracked along with his backpack. A blue square is drawn around the backpack. (b) The internal representation of each blob in the backpack module. The dark grey vertical line represents the computed symmetry axis, the white regions are considered symmetric, and the grey regions are considered asymmetric.

- For each human blob B'_p :
 - Scale B_p to 9x15 pixels and back to the original size to obscure details
 - * If B_p has not been seen before, store the location of its two left-side corners and centroid.
 - If B_p has been seen before, warp it using an affine transformation such that its two left-side corners and centroid are in the same locations as the first time the blob was seen. This aligns the blob in each subsequent frame.
 - * Compute the blob's symmetry axis V (vertical line in Fig. 5b) as the maximum value of the vertical projection, where y_0, y_1 are the left and right bounds of the blob, respectively:

$$V(x) = \int_{y_0}^{y_1} B(x, y) dx \quad (9)$$

- * Remove portions of which are reflections across the blob's symmetry axis within a margin of error, E (white in Fig. 5b). We use $E = 5\%$.
- * Store the remainder of the silhouette, i.e. the asymmetric region (grey in Fig. 5b), call it S_p .
- * Compute and store the similarity M_p between S_p and every previously stored region S'_p using (10), noting that S_p is a binary image:

$$M_p = \int_{x_0}^{x_1} \int_{y_0}^{y_1} S_p(x, y) - S'_p(x, y) dy dx \quad (10)$$

- * Compute the mean power spectrum P of all the stored M'_p values for this silhouette.
- * Where D is a scalar, if there is a peak in P :

$$P(x) > \text{mean}(P) + D * \text{std_dev}(P) \quad (11)$$

then mark S_p is asymmetric, aperiodic and believed to be a backpack. We use $D = 5$, chosen to scale $\text{std_dev}(P)$ to only mark peaks varying more than 99.5% from the mean.

Fig. 6: The Backpack Detection Algorithm

5. Experimental Evaluation

To validate our approach we collected sample sequences of people walking around, interacting with each other, and interacting with baggage. Two examples of each event type were recorded. In each pair of sequences the environments differ in room color/design, lighting, camera position, and subjects' clothing color, in order to demonstrate the system robustness to various environments. Each video is manually tagged to determine when each event instance begins and ends, and this is used as the ground truth for our validation. The system is then evaluated on three metrics from [19]:

- *Accuracy* = the percentage of observation sequences in which the system's selected event matches the ground truth.
- *Early detection* = $S' - S$, where S is the start time of the event according to the ground truth and S' is the start time detected by the system, in frames. Note that negative values are possible if the system detects the event before the ground truth.
- *Correct duration* = $\frac{T'}{T}$ where T is the duration of the event according to the ground truth and T' is the duration of the event detected by the system. Note that values greater than 1 are possible if the system detects the event longer than the ground truth.

For reliable event detection, the system should have high accuracy, a small value (close to zero) for *early detection*, and a large percentage (close to 1) for correct duration. The accuracy rate of our system is 100% in every sequence tested. Note that redundant events are not reported, e.g. an instance of two people meeting is also two separated instances of a single person. Table 1 shows the values for *early detection* and correct duration for each sequence.

For 10 of 12 cases the system detected the event within 16 frames (approximately one second) of the ground truth start time. The backpack module takes longer to begin tracking an event by design: periodicity is a function of time, thus strong peaks in the power spectrum of the backpack Fourier analysis only begin to appear about 60 frames (approximately two seconds) after the object tracking begins.

Event start and end times for the ground truth were taken to occur when a person made a visible attempt to commit the action, for instance reaching out to take the bag in the bag exchange and steal events. The variance for *early detection* is primarily due to these actions not necessarily corresponding to the method for determining events as described in Section 4. For event end times (which affect *correct duration*) there is variance among different interaction types. For instance, in BE2 the system detected a duration almost two times longer than the one indicated in the ground truth. This was observed because the people in the sequence walked past each other after the bag exchange, thus they were still in close proximity to each other and the bag, even though the

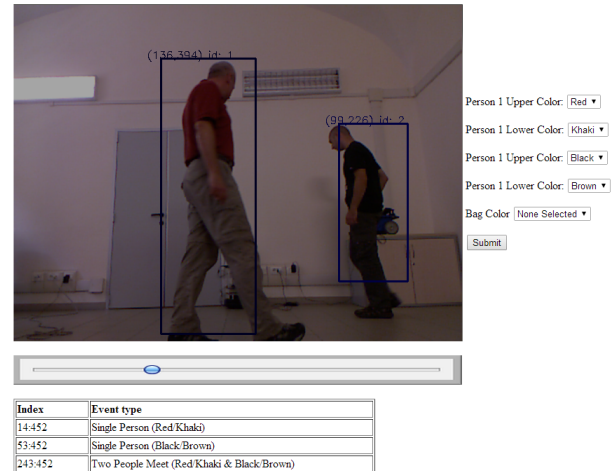


Fig. 7: The offline interface. A recorded video is shown with overlays, the user selects an option from the menu on the right of the video and clickable events are listed below the video.

Table 1: Quantitative Evaluation

Event	Early Detection (frames)	Correct Duration
SP1	3	1.0000
SP2	4	0.9591
TP1	2	1.1210
TP2	16	0.9012
BA1	5	0.9896
BA2	5	0.9829
BE1	-1	1.1905
BE2	2	1.7407
BS1	-8	1.0096
BS2	2	0.9551
PK1	56	0.6419
PK2	61	0.6324

exchange had finished.

We believe these results accomplish the objective of correctly identifying events to draw a human operator's attention. Slight variances in start and end times are insignificant in this case because an operator could simply rewind the video and play back as much of the sequence as they desire. There is an ambiguity in some event classes which was removed before determining a ground truth for some events. For a single person, two people meeting, and bag unattended it is easy to determine when the events start and end. The other events are more difficult to codify: does a bag theft happen when the owner initially leaves the bag unattended? Or when the thief approaches the bag? Or when the thief actually moves the bag? To disambiguate these circumstances the following rules are applied:

- A bag theft begins when a thief moves a bag from the location in which it was abandoned, and ends when the thief replaces the bag or the thief is no longer visible.

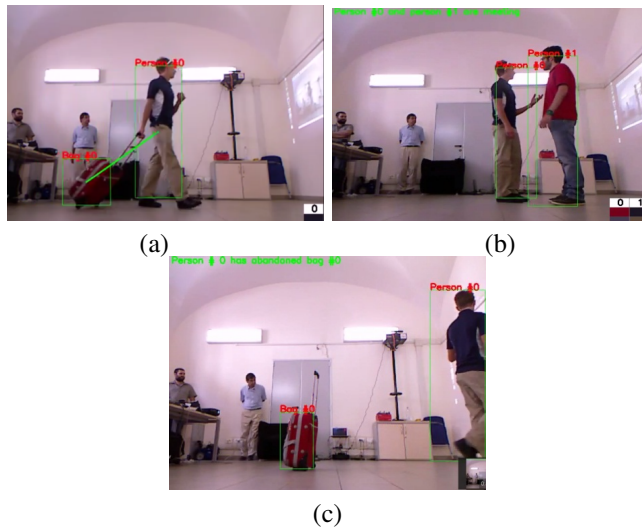


Fig. 8: Various events shown on the interface. (a) A single person is tracked along with his bag. A line connects the person and bag to express ownership. The people in the background are not highlighted because they do not match the description provided by the user. (b) A meeting between two people is detected and the system is providing an alert (top of the image). The top half of one person is largely occluded by his arm, but the system is still able to identify him. (c) A single person is tracked along with his bag. The owner has walked away from his bag and the system is providing an alert (top of the image). No line is drawn between the person and bag, showing no ownership.

- A bag exchange begins when the original bag owner and bag receiver are close to the bag and ends when the original owner is no longer close to the bag or the receiver is no longer visible. Note that the system may begin detecting a bag exchange and then the “receiver” may leave without taking the bag (this would be correctly classified as two people meeting). In this case, our system deletes the false event instance.

6. Conclusion and Future Work

In this paper we presented a robust system for detecting human agents and interactions with the objects they carry. We use computer vision techniques to segment and track people and bags, and infer the relationships between them. We also use periodicity and symmetry information to perform a Fourier analysis on human blobs to determine if the person is carrying a backpack or other large item. The proposed system demonstrates the potential of fully automated, vision-based solutions for detecting predefined behaviors and reporting instances of these events to a human operator. Such system would be of use to security personnel in sensitive public places such as airports or federal buildings. The system would significantly reduce the workload of human operators required to manually detect security events from live video streams. Additionally, the offline mode allows security or law enforcement personnel to search in recorded video for specific events much faster than replaying and watching an entire video.

We plan to expand this system by adding an autonomous robotic element to the detection system in order to provide information that cannot be obtained by a static camera. For instance, a robot could be deployed to a certain location to get a better view of a particular event, resolve occlusions, or provide a more detailed view of a person involved in a specific action.

References

- [1] A. Fong and S. Hui, “Web-based intelligent surveillance system for detection of criminal activities,” *Comput. & Control Eng. J.*, vol. 12, pp. 263-270, 2001.
- [2] J. S. Kim, D. H. Yeom, Y. H. Joo, and J. B. Park, “Intelligent unmanned anti-theft system using network camera,” *Int. J. Control, Automation, Syst.*, vol. 8, pp. 967-974, 2010.
- [3] S. J. Junior, “Crowd analysis using computer vision techniques,” *IEEE Signal Process. Mag.*, vol. 27, pp. 66-77, 2010.
- [4] T. D. Raty, “Survey on contemporary remote surveillance systems for public safety,” *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 40, pp. 493-515, 2010.
- [5] M. Valera and S. Velastin, “Intelligent distributed surveillance systems: a review,” in *Proc. 2005 IEE Vision, Image and Signal Process.*, 2005, pp. 192-204.
- [6] D. H. Harris, “How to really improve airport security,” *Ergonomics in Design: The Quart. of Human Factors Applicat.*, vol. 10, pp. 17-22, 2002.
- [7] L. Iocchi, D. Monekosso, D. Nardi, M. Nicolescu, P. Remagnino, and M. Valera, “Smart monitoring of complex public scenes,” *AAAI Fall Symp.*, Arlington, VA, 2011, pp. 14-19.
- [8] L. Rothkrantz, “Crisis Management Using Multiple Camera Surveillance Systems,” *Int. Inform. Syst. for Crisis Response and Manage.*, Baden-Baden, 2013, pp. 617-626.
- [9] M. Kylanpaa, A. Rantala, J. Merilinna, and M. Nieminen, “Secure communication platform for distributed city-wide surveillance systems,” in *4th Int. Conf. Inform., Intell., Syst. and Applicat.*, 2013, pp. 1-4.
- [10] I. Thibault, “Advanced Beamforming for Distributed and Multi-Beam Networks,” Ph.D. dissertation, Dept. of Elect., Electron., and Inform. Eng., Univ. of Bologna, Bologna, 2013.
- [11] A. Hilal, “An Intelligent Sensor Management Framework for Pervasive Surveillance,” Ph.D. dissertation, Dept. of Elect. and Comput. Eng., Univ. of Waterloo, Waterloo, ON, 2013.
- [12] I. Haritaoglu, D. Harwood, and L. S. Davis, “W4: Real-time surveillance of people and their activities,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 809-830, 2000.
- [13] T. Zhao, R. Nevatia, and F. Lv, “Segmentation and tracking of multiple humans in complex situations,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition*, 2001, pp. II194-II201 vol. 2.
- [14] R. Cutler and L. Davis, “View-based detection and analysis of periodic motion,” in *Int. Conf. Pattern Recognition*, 1998, pp. 495-495.
- [15] Z. Zivkovic and F. van der Heijden, “Efficient adaptive density estimation per image pixel for the task of background subtraction,” *Pattern recognition letters*, vol. 27, pp. 773-780, 2006.
- [16] S. Suzuki, “Topological structural analysis of digitized binary images by border following,” *Comput. Vision, Graph., and Image Process.*, vol. 30, pp. 32-46, 1985.
- [17] R. A. Fisher, “Tests of Goodness of Fit, Independence and Homogeneity” in *Statistical Methods for Research Workers*. Edinburgh, Scotland: Oliver and Boyd, 1925, ch. IV.
- [18] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *J. Basic Eng.*, vol. 82, pp. 35-45, 1960.
- [19] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, “Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model,” in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition*, 2005, pp. 955-960.

Applying A Lightweight Chinese Lexical Chain Processing In Web Image Annotation

Chuen-Min Huang

Department of Information Management

National Yunlin University of Science & Technology, Taiwan, R.O.C.

huangcm@yuntech.edu.tw

Tel: +886-9-963119624

Abstract - Traditional CBIR method relies on visual features to identify objects in an image and uses predefined terms to annotate images, thus it fails to depict the implicit meanings. Recent textual-based analysis methods only focus on single term processing thereafter they suffered the disadvantages of fragmented description of the annotation. In this research, we propose a corpus-free, relatively light computation of term segmentation method, namely 'Chinese Lexical Chain Processing (CLCP),' to identify compounds from a single web page to obtain anecdotes as a semantic enrichment of the target image. It requires a minimum computation need that allows sharing characters/words and facilitating their use at fine granularities without prohibitive cost. In the experiment, this method achieves a precision rate of 90.04%, and gains acceptance from expert rating and user rating of 86% and 73.7%, respectively. In performance testing, it only takes 0.007 second to process each image in a collection of 1,728 testing data set.

Keywords: Automatic Image Annotation, Chinese Lexical Chain Processing, N-Gram, Lexical Chain

1 Introduction

With the rapid development of image technology and digital devices, more and more web images are stored and displayed on the Internet. As a result, the way to process these images to be retrieved efficiently has become a crucial issue. Content-Based Image Retrieval (CBIR) with its focus on rapid application of voluminous low-level visual features such as color, texture, shape, etc. gains popularity to support efficient searching and browsing images. Due to the visual features explain less semantics, and the annotation relies on limited predefined terms, thereby the results usually are not satisfactory.

A glimpse of related studies would reveal that a couple of supervised learning approach and clustering techniques have been applied to image annotation including Support Vector Machine (SVM) [1, 2], Bayesian [3] and Self-Organizing Feature Map (SOM) [4, 5]. In addition, various text processing techniques that support content identification to analyze textual content based on word co-occurrence, location, and lexical-chained concepts have been elaborated in [6-8]. However, the aforementioned techniques suffer the same disadvantages of heavy computation for multi-document processing, which will consume lots of memory and may incur run-time overhead. In this study, we propose a corpus-free, relatively light computation of term segmentation for single document processing, namely 'Chinese Lexical Chain Processing (CLCP) method.' The CLCP method is to identify representative terms from a string in a single document with minimum computation needs that allows sharing characters/words and facilitating their use at fine granularities without prohibitive cost. The results demonstrated that

applying our method in a single document is enough to generate content descriptor for image annotation. The precision rate achieves 90.04%, and the acceptance from expert and user rating reach 86% and 73.7%, respectively. The performance testing is also very promising.

The remainder of the paper is organized as follows. Section 2 presents the related work including automatic image annotation, keyword extraction, and lexical chain. Then, we address the intent of our experiment in Section 3. The experimental result and evaluation method are described in Section 4. Finally, in Section 5, we draw conclusions and suggest future work.

2 Literature Background

2.1 Automatic Image Annotation

There have been a number of models applied for image annotation. In general, image annotation can be categorized into three types: retrieval-based, classification-based, and probabilistic-based [6]. The basic notion behind retrieval-based annotation is that semantic-relevant images are composed of similar visual features. CBIR have been proposed in 1992 [9]. Since then, more and more studies annotated the images based on this method [10]. CBIR is applied by the use of images features, including shape, color and texture. However, this method is limited by the training data set and the hidden semantic or abstract concepts can't be extracted because the keywords are confined to pre-defined terms. Consequently, the results of CBIR are usually not satisfactory. The second type, also known as the supervised learning approach, treats annotation as classification using multiple classifiers. The images are classified based on the

extracted features. This method processes each semantic concept as an independent class, and assigns each concept as one classifier. Bayesian [3] and SVM [11] are the most often used approaches. The third type is constructed by estimating the correlations between images and concepts with a particular emphasis on the term-term relationship and intends to solve the problem of 'synonym' and 'homograph.' Frequent used approaches include co-occurrence model [12], LSA [5], PLSA [13] and HMM [14]. Notwithstanding the efforts made on the enhancement of annotation quality, the aforementioned approaches only focused on single term processing thereafter they suffered the disadvantages of fragmented description of annotation.

2.2 Keyword Extraction

In the field of Information Retrieval, keyword extraction plays a key role in summarization, text clustering/classification, and so on. It aims at extracting keywords that represents the text theme. One of the most prominent problems in processing Chinese texts is the identification of valid words in a sentence, since there are no delimiters to separate words from characters in a sentence. Therefore, identifying words/phrases is difficult because of segmentation ambiguities and the occurrences of newly formed words. In general, Chinese texts can be parsed using dictionary lookup, statistical or hybrid approaches [15].

The dictionary lookup approach identifies keywords of a string by mapping well-established corpus. For example, the Chinese string '蔡英文北監探扁會面一小時' (Ying-wen Tsai went to Taipei prison to visit Shui-bian Chen and talk for an hour) will be parsed as: '[蔡英文],[北監],[探],[扁],[會面],[一小時]' by a well-known dictionary-based CKIP segmentation system in Taiwan. This method is very efficient while it fails to identify newly formed or out-of-the-vocabulary words and it is also blamed for the triviality of the list of the extracted words.

The statistical technique extracts elements by using n-gram (bi-gram, tri-gram, ... etc.) computation from the input string. This method relies on the frequency of each n-gram and a threshold to determine the validity of each word. The above string through n-gram segmentation will produce: '[蔡英文],[英文],[文北],[北監],[監探],[探扁],[扁會],[會面],[面一],[一小時]'; '[蔡英文],[英文北], ..., [一小時]' and so on. The application of this method has the benefit of corpus-free and the capability of extracting newly formed or out-of-the-vocabulary words while at the expense of huge computations and the follow-up filtering efforts.

Recently, a number of studies proposed substring [9], significant estimation [16], and relational normalization [17, 18] to identify words based on statistical calculations. The hybrid method conducts dictionary mapping to process the major task of word extraction and handle the leftovers through n-gram computation, which significantly reduces the amount of terms under processing and takes care both the quality of term segmentation and the identification of unknown words. It has gained popularity and adopted by many researchers [19, 20]. Since the most important task of

annotation is to identify the most informative parts of a text comparatively with the rest. Consequently, a good text segmentation shall help in this identification.

In the IR theory, the representation of documents is based on the Vector Space Model[21]: a document is a vector of weighted words belonging to a vocabulary V : $d = \{w_1, \dots, w_{|V|}\}$. Each weight w_n is such that $0 \leq w_n \leq 1$ and represents how much the term t_n contributes to the semantics of the document d . In the term frequency-inverse document frequency (tf-idf) model, the weight is typically proportional to the term frequency and inversely proportional to the frequency and length of the documents containing the term. The term discrimination value can be used to compute a weight for each word in each document of a collection by combining the term frequency factor with the discrimination value. Some studies [22] [23] proposed methods to assign different weights to words by location.

2.3 Lexical Chain

A lexical chain (LC) is a sequence of words, which is independent of the grammatical structure of the text. Lexical cohesion can be interpreted as the state of cohering for making the sentences of a text, indicated by the use of semantically related vocabulary. Lexical chains (LCs) are sequences of words which are in lexical cohesion relations with each other and they tend to indicate portions of the context that form semantic units; they could serve further as a basis for a segmentation [24]. This method is usually applied in a summarization generation [25]. For instance, the string '向量空間模型' (Vector space model) may be parsed as '[向量],[空間],[模型]' if there is no further merging process undergoing. Thereby the most informative compound '向量空間模型' will be left out. Usually LCs are constructed in a bottom-up manner by taking each candidate word of a text, and finding an appropriate semantic relation offered by a thesaurus. Instead, the paper [26] proposes a top-down approach of linear text segmentation based on lexical cohesion of a text. Some scholars suggested to use machine learning approaches to create a set of rules to calculate the rate of forming a new word by characters for entity recognition including the maximum entropy model (MEM) and hidden Markov model (HMM) and claimed this method was able to achieve reasonable performance with minimal training data [27, 28].

3 Research Design

3.1 Research Model Overview

This study covers three tasks: text processing, image annotation, and image evaluation. The framework of our research is depicted as **Figure 1**. We used the news title and the text as input data; the image captions as ground truth labels. Then, we conducted CLCP and term weighting for the input data. After that, we generated a list of three

representative words for each image. From the list, we assigned the word with the highest weight as the primary annotation and the rest as secondary annotations. Finally, we evaluated the primary and secondary annotations by using the image caption and human judgment, respectively. In the following section, we will introduce the process of CLCP and the way of term weighting and the word annotation.

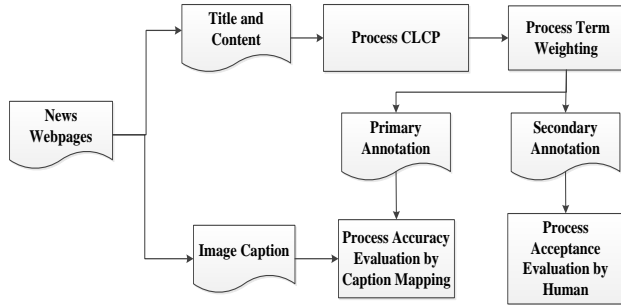


Figure 1. Research model

3.2 CLCP

The fundamental idea of building CLCP was a bottom-up concatenating process based on the significance degree of distribution rate to extract the most meaningful LCs in a string. We treated a news document as a string composed of a series of characters and punctuations. Since there is no delimiter to separate words from characters except for the usage of quotation marks in special occasion, this concatenating process is a challenge without the aid of dictionary.

Most of the traditional studies identify words from the whole context and store all the keywords for further processing as **Figure 2**. In this way, even a moderate-sized document may require hundreds of thousands of characters, which will consume lots of memory and may incur unacceptable run-time overhead. Due to the number of distinct characters processed is less than that in the document, we adopted a sharing concept to allow reuse of the identical characters. We considered each character as a basic unit from which to build compounds as a composite, which in turn can be grouped to form larger compounds. Since the character and compound will be treated uniformly, it makes the application simple.

By doing so, we adopted the concept of flyweight and composite design patterns proposed by the GOF [29] to implement this design. **Figure 3** shows the flyweight as a shared object that can be used in the whole context simultaneously. **Figure 4** represents the part-whole hierarchy of texts and the way to use recursive composition. By applying flyweight design pattern, it supports the use of large numbers of fine-grained objects efficiently. By applying composite design pattern, it makes the application easier to add new components. The CLCP steps are depicted as

Figure 5. The detailed description with an example is addressed next.

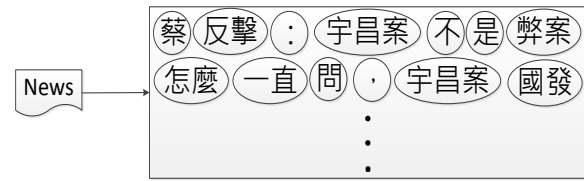


Figure 2 Traditional document processing

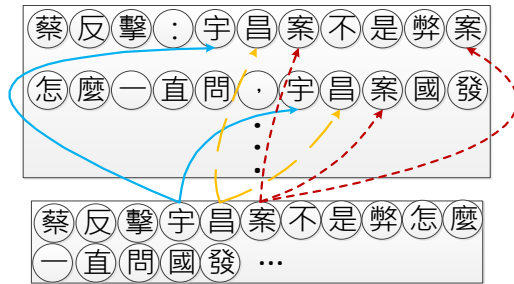


Figure 3 Flyweight design concept

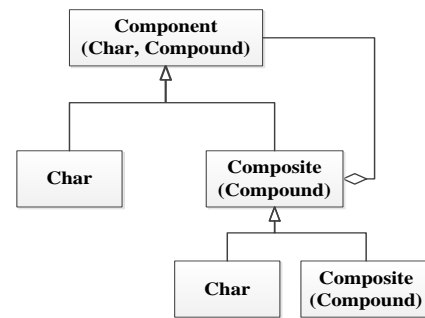


Figure 4 Composite text structure

3.2.1 Step 1: Build a directed graph

A directed graph (or digraph) is a set of nodes connected by edges, where the edges have a direction associated with them. For example, an arc (x, y) is considered to be directed from x to y , and the arc (y, x) is the inverted digraph. y is the head and x is the tail of the link; y is a direct successor of x , and x is a direct predecessor of y .

We use the string: '蔡反擊：宇昌案不是弊案，怎麼一直問，宇昌案國發基金為何一再放棄權利...' (Tsai fired back and stated that Yu Chang case is not a scandal, why did you keep asking for this and why the National Development Fund abandoned its rights repeatedly?) as an example to explain the construction process. After removing duplicate characters and replacing punctuations with new lines, a digraph is built. **Figure 6** presents a fraction of the graph in which a solid line indicates the directed link and a dash line means the inverted link.

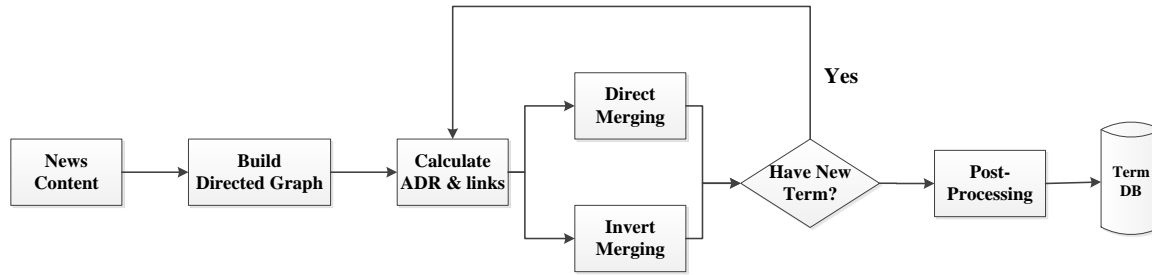


Figure 5 CLCP steps

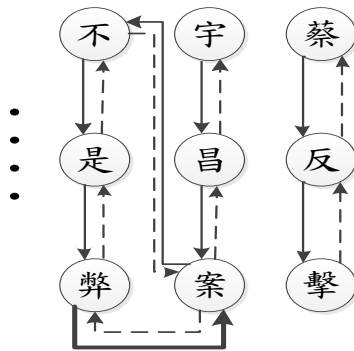


Figure 6. A fraction of directed graph

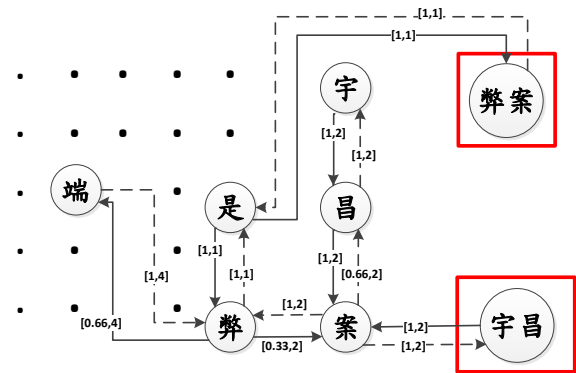


Figure 9. Concatenate vertices

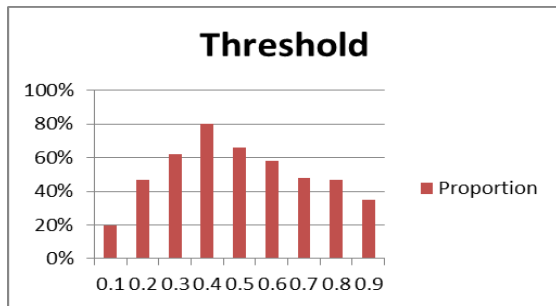


Figure 7 The threshold scatter diagram

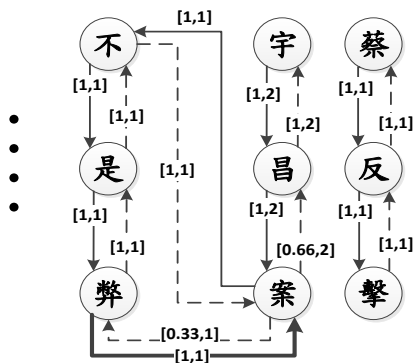


Figure 8. The fraction of ADR

3.2.2 Step 2: Calculate ADR and concatenate vertices

This step is to calculate the intensity and the degree of a digraph. The intensity is the average distribution rate (ADR); the degree means the number of incident edges. To determine whether two vertices can be concatenated, we applied the criteria listed in (1), (2).

$$D(i,j) \geq T \cap (\text{Digraph}(\text{Node}_i) > 1) \tag{1}$$

$$D(p,q) \geq T \cap (\text{InvD}(\text{Node}_p) > 1) \tag{2}$$

Where $D(i,j)$ and $D(p,q)$ represent the ADR of the arcs (i, j) and (p, q) ; $\text{Digraph}(\text{Node}_i)$ and $\text{InvD}(\text{Node}_p)$ indicate the number of directed links for vertex i and inverted links for vertex p , respectively. T is the threshold value determined by 10 runs of experiments with the value of 0.1~0.9 assigned to 100 documents and the result was verified by 5 subject specialists with respect to word quality. The result showed that 0.4 outperforms the others as **Figure 7**, thus we used it as the threshold value. Based on the term frequency theory, it infers that the significance of the concatenation will be proportional to its frequency. Since this study was to generate content descriptor from a single document, the frequency was set as reasonable as possible.

Figure 8 shows the arc [宇, 昌] with the expression [1,2] indicating that the intensity is 1 and the degree is 2, therefore it will be concatenated as [宇昌] because it meets the criteria. In our previous [30], we focused mainly on directed links to extract LCs and failed in identifying some significant concatenations when a vertex has many links which diluted

the intensity of the distribution rate. This study was intended to remedy this problem by processing the directed link and the inverted link in parallel. In **Figure 9**, the LC [弊案] with the value [0.33, 2] of directed link [弊, 案] will be left out if we didn't consider its inverted link [案, 弊] with the value [1,2].

3.2.3 Step 3: Run Iteration

The above steps will be iterated until no concatenation can be found, and this iteration process will generate a series of short and long LCs from the string. A long LC is believed to be more content representative than a short LC could possibly be. In this example, it is obvious that '字昌案' is a better content indicator than either '字昌' or '昌案.' To reduce the possibility of extracting less representative LCs from the concatenation, we will take a post-processing as the last step to finalize the CLCP.

3.2.4 Step 4: Execute Post-processing

In the final step, significant words are determined by observing the information mutually shared by two-overlapped LCs using the following significance estimation (SE) function as (3).

$$SE_i = \frac{f_i}{f_a + f_b - f_i} \quad (3)$$

Where i denotes the LC_i to be estimated, i.e., $i = i_1 i_2 \dots i_n$; a and b represent the two longest compound substrings of LC_i with the length $n-1$, i.e., $a = i_1 i_2 \dots i_{n-1}$ and $b = i_2 i_3 \dots i_n$. As for f_a , f_b and f_i are the frequencies of a , b , and i , respectively. For example, the term i , '字昌案', shall gain the SE value of 0.83 with its frequency 5 and the frequency 6 of its substring a , '字昌', as well as the value 5 of the other substring '昌案'. In this case, we will retain term i '字昌案' and its substring a '字昌' because the frequency of '字昌案' is less than '字昌' indicating '字昌' implies useful meanings. Likewise, we will discard the substring b '昌案' because both terms have the same frequency indicating the long term '字昌案' can replace its substring '昌案.' As stated above, since $f_i < f_a$, we retain both terms, and discard '昌案' because $f_i = f_b$.

3.3 Term Weighting

It is suggested that the obvious place where appropriate content identifiers might be found in news is the title and the first paragraph. In addition, we also considered frequency and the length of a word as the indicators of word significance in a document. Given a word LC_i , the term weighting algorithm may be defined as (4).

Where tf_i represents the frequency of LC_i ; val_1 and val_2 express the extra weight of LC_i based on its position in the news report. If LC_i appears in the title or the first paragraph, it gains an extra weight of 2 respectively.

$$Weight_i = tf_i \times (val_1 + val_2 + length_i)$$

$$val_1 = \begin{cases} 2, & word_i \in title \\ 0, & otherwise \end{cases} \quad val_2 = \begin{cases} 2, & word_i \in FP \\ 0, & otherwise \end{cases} \quad (4)$$

4. Evaluation

In this experiment, we collected 1,738 images-resided web pages from Taiwan news website udn.com as the data sets. To verify our proposed CLCP method can successfully identify the content representation for image annotation, we used image captions as ground truth labels to see whether the primary annotation is included in the list. Subject experts and users were invited to assess the appropriateness of the secondary annotations (rank the second and the third). In the end, we also measured the performance of the CLCP method in a real-time mode.

4.1 Evaluation of Primary Annotation

Since the image captions are written by journalists, it is assumed that a man-made caption would be faithful to an image scenario. Therefore, we considered image captions as the ground truth labels, which will be used to evaluate the accuracy rate of the produced image annotations from the title and text. If the primary annotation matches a substring of the image caption, we will be confident to assure that the CLCP method works satisfactory.

Due to the problem of semantic ambiguity in part of Chinese words, where the interpretation of image annotations may vary from users to users, therefore the exact number of correct annotations of an image will not be clearly identified. For example, the string '總統馬英九' (President Ma Ying-jiu) is meant to be regarded as a single LC, therefore it may not be appropriate to segment it into [總統] and [馬英九] even though these two words are valid. Thus, the recall measurement did not apply to this study. A precision measurement was used to understand the proportion of primary annotations actually matched the image captions as (5).

$$p = \frac{\text{number of matched PAs in captions}}{\text{total number of PAs}} \quad (5)$$

Where PAs represent the generated primary annotations from 1,738 documents. After the CLCP processing, the total number of matched PAs in captions are 1,565. We obtained a precision rate of 90.04%.

4.2 Evaluation of Secondary Annotation by Expert

To evaluate the validity of the secondary annotations, we invited five subject experts to participate in the assessment. Thirty pieces of news were randomly selected from which we extracted second and third place in scores of LCs and produced 60 annotations. To reduce ambiguous judgements, each annotation was evaluated based on a method of dichotomic classification to which the annotation represents

the image content. Each expert could only select either 'agree' or 'disagree' for each annotation. The result showed that the number of check marks of consent is 310 out of 360. It implies that the agreement of the appropriateness of the secondary annotations to the images achieves 86%.

4.3 Evaluation of Secondary Annotation by User

To evaluate the appropriateness of the secondary annotation, we conducted another survey to understand the differences between the image annotation and the users' expectation. Sixty graduate and undergraduate students were recruited from National Yunlin university of Science & Technology, Taiwan to participate in the assessment. Sixty pieces of news were randomly selected from which we extracted second and third place in scores of LCs and produced 120 annotations. To assist the assessment, we provided news title, texts and caption for references. Each annotation was evaluated by these participants to understand the degree to which the annotations appropriately address the image content. The result in Table 1 shows that the number of check marks of agreement is much higher than that of disagreement. The agreement rate of user evaluation reaches 73.7%.

Table 1. Statistics of results of user satisfaction

<i>Highly Agree</i>	<i>Agree</i>	<i>Average</i>	<i>Disagree</i>	<i>Highly Disagree</i>	<i>Total</i>
1620	1238	684	218	116	3876

4.4 Performance Testing

After the validity and acceptance evaluation, we conducted a performance testing with respect to the time spent of processing from an event trigger to system response. Often real-time response times are understood to be in milliseconds and sometimes microseconds. Our testing data sets consist of 1,738 pieces of news; the processing time is 12.52 seconds in total with 0.007 seconds on average for each piece of news.

5. Conclusion

In this paper, we propose a corpus-free, relatively light computation of term segmentation for single document processing, namely "Chinese Lexical Chain Processing (CLCP) method" to identify representative terms for image annotation. The CLCP method is based on a hybrid of n-gram and lexical chain processing for image annotation. Unlike recent textual-based analysis methods only focused on single term processing thereafter they suffered the disadvantages of fragmented description of the annotation. We considered each character as a basic unit from which to build compounds as a composite, which in turn can be grouped to form larger compounds. Since the character and compound will be treated uniformly, it makes the application simple.

Our method allows sharing characters/words and facilitating their use at fine granularities without prohibitive cost. Results showed that this method achieves a precision rate of 90.04%, and gains acceptance from expert and user rating of 86% and 73.7%, respectively. In performance testing, it only takes

0.007 second to process each image in a collection of 1,728 testing data set.

Even though the CLCP method is only applied to single-document processing in the current study, with the virtue of corpus-free and lightweight features, it shall gain more benefits from applying in multi-document processing. Other researchers may verify our study using a larger data set or compare with the state of the art algorithms. It is hoped that our research will invite more perspectives on automatic image annotation.

6. Acknowledgements

This study was supported by the National Science Council, Taiwan, Republic of China, under the project of special research projects of free software (NSC101-2221-E-224-056).

References

- [1] S. Gao, D.-H. Wang, and C.-H. Lee, "Automatic image annotation through multi-topic text categorization," presented at the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006.
- [2] Z. Lei and M. Jun, "Image annotation by incorporating word correlations into multi-class SVM," *Soft Computing*, vol. 15, pp. 917-927, 2011.
- [3] N. Luong-Dong, Y. Ghim-Eng, L. Ying, T. Ah-Hwee, C. Liang-Tien, and L. Joo-Hwee, "A Bayesian approach integrating regional and global features for image semantic learning," in *IEEE International Conference on Multimedia and Expo, 2009. ICME 2009*, 2009, pp. 546-549.
- [4] T. W. S. Chow and M. K. M. Rahman, "A new image classification technique using tree-structured regional features," *Advanced Neurocomputing Theory and Methodology*, vol. 70, pp. 1040-1050, 2007.
- [5] C.-M. Huang, C.-C. Chang, and C.-T. Chen, "Automatic image annotation by incorporating weighting strategy with CSOM classifier," presented at the The 2011 International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV'11), Monte Carlo Resort, Las Vegas, Nevada, USA, 2011.
- [6] J. H. Su, C. L. Chou, C. Y. Lin, and V. S. Tseng, "Effective image semantic annotation by discovering visual-concept associations from image-concept distribution model," *IEEE International Conference On Multimedia. Proceedings*, pp. 42-47, 2010.
- [7] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *The Journal of Machine Learning Research*, vol. 3, pp. 1107-1135, 2003.
- [8] S. Zhu and Y. Liu, "Semi-supervised learning model based efficient image annotation," *Signal Processing Letter, IEEE*, vol. 16, pp. 989-992, 2009.

- [9] T. Kato, "Database architecture for content-based image retrieval," in *Proc. SPIE 1662, Image Storage and Retrieval Systems*, 1992, pp. 112-123.
- [10] L. Jing, M. Shao-ping, and Z. Min, "Automatic image annotation based-on model space," in *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering 2005*, pp. 455-460.
- [11] Y. Gao, J. Fan, X. Xue, and R. Jain, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers," presented at the Proceedings of the 14th Annual ACM International Conference on Multimedia, Santa Barbara, CA, USA, 2006.
- [12] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," presented at the First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.
- [13] F. Monay and D. Gatica-Perez, "PLSA-based image auto-annotation: constraining the latent space," presented at the Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004.
- [14] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
- [15] Z. Wang, J. Xu, A. K., and T. K., "Word segmentation of Chinese text with multiple hybrid methods," presented at the 2009 International Conference on Computational Intelligence and Software Engineering, 2009.
- [16] J. T. Horng and C. C. Yeh, "Applying genetic algorithms to query optimization in document retrieval," *Information Processing & Management*, vol. 36, pp. 737-759, 2000.
- [17] M. S. Kim, K. Y. Whang, J. G. Lee, and M. J. Lee, "Structural optimization of a full-text n-gram index using relational normalization," *The VLDB Journal*, vol. 17, pp. 1485-1507, 2008.
- [18] M. Fuketa, N. Fujisawa, H. Bando, K. Morita, and J. i. Aoe, "A retrieval method of similar strings using substrings," presented at the 2010 Second International Conference on Computer Engineering and Applications, 2010.
- [19] C.-M. Hong, C.-M. Chen, and C.-Y. Chiu, "Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems," *Expert Systems with Applications*, vol. 36, pp. 3641-3651, 2009.
- [20] R. T.-H. Tsai, "Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures," *Expert Systems with Applications*, vol. 37, pp. 3553-3560, 2010.
- [21] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company, 1983.
- [22] H. Yan, S. Ding, and T. Suel, "Compressing term positions in web indexes," presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, MA, USA, 2009.
- [23] A. D. Troy and G.-Q. Zhang, "Enhancing relevance scoring with chronological term rank," presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007.
- [24] D. Tatar, E. Kapetanios, C. Sacarea, and D. Tanase, "Text segments as constrained formal concepts," in *12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2010* 2010, pp. 223-228.
- [25] V. Shanthi and S. Lalitha, "Lexical chaining process for text generations," presented at the International Conference on Process Automation, Control and Computing (PACC), 2011.
- [26] D. Tatar, A. D. Mihis, and G. S. Czibula, "Lexical chains segmentation in summarization," in *10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2008, pp. 95-101.
- [27] R. Chiong and W. Wang, "Named entity recognition using hybrid machine learning approach," presented at the The 5th IEEE International Conference on Cognitive Informatics 2006.
- [28] R. Ageishi and T. Miura, "Named entity recognition based on a Hidden Markov Model in part-of-speech tagging," presented at the First International Conference on the Applications of Digital Information and Web Technologies, 2008.
- [29] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Indianapolis, IN: Pearson Education Co., 1995.
- [30] C.-M. Huang and Y.-J. Chang, "Applying a lightweight iterative merging Chinese segmentation in web image annotation," in *International Conference on Machine Learning and Data Mining* New York, USA, 2013, pp. 183-194

A Heterogeneous Fuzzy Clustering Approach for Reliable Audio Genre Classification

Junsang Seo¹, Myeongsu Kang¹, Cheol-Hong Kim², and Jong-Myon Kim^{1,*}

¹Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, South

²School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea
 jsseo2006@gmail.com, ilmareboy@ulsan.ac.kr, chkim22@chonnam.ac.kr, jmkim07@ulsan.ac.kr

Abstract - To discriminate different audio genres in an unknown incoming audio stream, this paper proposes a heterogeneous fuzzy clustering scheme, which unifies objective functions of the following fuzzy clustering methods: fuzzy *c*-means (FCM), possibilistic *c*-means (PCM), and hard *c*-means (HCM). The proposed hybrid fuzzy clustering methodology utilizes two parameters ranging from 0 to 1 in order to control the amount of contributions of the objective function of FCM, PCM, and HCM. This results in enhancement of the proposed scheme for audio classification in terms of the consistency and computational efficiency compared to the conventional FCM-based approach. More specifically, the proposed approach reduces the effect of sensitivity to outliers and convergence speed. To show the improved performance of the proposed approach, this paper first extracts audio features in time and frequency domains, and such audio features are then utilized as inputs of the proposed heterogeneous fuzzy clustering approach. Experimental results indicate that the proposed approach outperforms the conventional FCM-based audio classification approach with regard to classification accuracy (i.e., 46.99% improvement in accuracy) and computational complexity (i.e., 40.78% improvement in execution time), respectively

Keywords: Audio classification, audio content retrieval, fuzzy clustering, multimedia indexing, pattern recognition

1 Introduction

Rapid increase of multimedia contents in computer applications build a large number of audio databases having vast amount of audio signals at user-ends, which consist of different primitive audio genres such as speech (SP), music (MS), speech with music (SM), speech with noise (SN), and silence (SP) [1]. Classification of audio into these genres can be considered the primary step before further audio processing. Now-a-day, a traditional classification method is being replaced by the automated computerized content-based

approach that can be divided into two tasks: feature extraction and classification [2, 3]. Essentially, a carefully selected efficient classification framework along with a competent feature vector can achieve the objective of efficient classification performance.

The area of content based audio classification was introduced by the Muscle Fish Group [4]. They used some acoustical features such as loudness, pitch, brightness, bandwidth, and the harmony of audio signals to discriminate some environmental sounds. Further, *Tzanetakis et al.* proposed a method by giving an insight into the nature of music signals, which afterward investigated by several music genres classification methods [2, 5]. On the other hand, *Saunders* made an attention to discriminate between broad audio classes like SP and MS [6]. Subsequently, a large number of methods have been reported to classify signals into primitive audio genres. For example, *Lu et al.* introduced a set of new audio features in both energy and frequency domains and proposed a segmentation algorithm that is based on quasi-Gaussian mixture model (GMM) and line spectral pair correlation analysis [7]. Support vector machine (SVM) classifier was introduced for audio classification by *Lu et al.* [8] and this approach was further improved by [9, 10]. However, these methods are subjected to the complexities of selection of kernel functions and computationally expensive training stages of multiclass SVMs. Recently, an efficient method was proposed by *Krishnamoorthy et al.* to select an optimal feature vector and they achieve considerably good performance in discriminating SP and MS [11].

Besides, the fuzzy *c*-means (FCM) algorithm has been frequently used in a number of works. For example, *Liu et al.* proposed a hierarchical fuzzy tree structure and utilized some time and frequency domain features to suit it with online web applications [12]. *Khan et al.* utilized FCM to select a viable set of features to get better classification accuracy [13]. *Nitanda et al.* proposed a new FCM-based method for segmentation of long audio signals and classified these into primitive audio genres [14]. They achieved good performance for segmentation; however the accuracy of classification was poor. Some derivatives of FCM have also been reported in this context. *Park et al.* proposed three different approaches for music genre classification utilizing the same feature vector of *Tzanetakis et al.* [5], and employing a gradient-

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. NRF-2013R1A2A2A05004566).

* Corresponding author.

based FCM algorithm (GBFCM) with divergence measures (GBFCM(DM)) [15], a GBFCM with mercer kernel (GBFCM(MK)) [16], and a FCM with divergence kernel (FCMDK) [17]. *Luong et al.* and *Nguyen et al.* improved the method of *Nitanda et al.* by use of two derivatives of FCM to reflect the temporal correlations of audio data in classification [18, 19].

FCM has been frequently used in audio classification and achieved considerable performance. This is due to the high partition quality of FCM and its easily understandable and implementable alternating optimization (AO) solution. However, all the above methods are subject to the facts that FCM is sensitive to outliers and the convergence speed of FCM is quite slow [20]. The sensitivity to outliers generates erroneous results and diminishes the robustness of the system. In addition, the slower convergence speed increases the computational cost. In order to solve these problems of FCM, a number of approaches have been proposed. A solution to the sensitivity to outliers was proposed by *Krishnapuram et al.* through a possibilistic c-means (PCM) algorithm which leads to a problem of coincident clusters as two or more prototypes can converge to the same position [21]. In order to solve this problem, *Pal et al.* proposed a possibilistic FCM (PFCM), by combining FCM with PCM [22]. They used a hybrid cost function devised from the objectives of FCM and PCM and utilized two tradeoff parameters to control the mixture of the objectives of FCM and PCM. On the other hand, *Fan et al.* proposed the suppressed FCM (SFCM) algorithm which utilized the hard c-means (HCM) algorithm in order to attain quicker convergence of FCM [23]. They added an extra computation step into the FCM iteration, which created a competition among clusters by suppressing the low membership values in FCM according to a previously defined suppression rate. This algorithm is able to reduce the computational cost of FCM by carefully choosing the suppression rate. *Szilagyi et al.* further improve SFCM by an optimal SFCM (OSFCM) [24].

The problems of FCM mentioned in [24] and the solutions proposed in previous works leads to the reexamination of FCM-based audio classification approaches. In order to improve the robustness, accuracy, and computational cost of the FCM-based approaches, it is necessary to combine the advantages of the work reported previously [21, 22] to address the problem of sensitivity to outliers and several methods are reported [23, 24] that solve the problem of slow convergence speed. In this paper, an audio classification framework is proposed that is based on a hybrid c-means clustering approach that includes the advantages of FCM, PCM and HCM in order to improve robustness, accuracy, and computational speed in audio classification. In addition, the k-nearest neighbor (kNN) algorithm [25] is utilized to automatically resolve the genre of clusters generated by the hybrid c-means approach. A competent feature vector was used along with the proposed classifier to classify audio signals into five broad genres such as SP, MS, SL, SN, and SM. For experimental evaluation, a

benchmark database that includes possible varieties of sounds was utilized and assessment of the performance of the proposed method was accomplished by comparing it with the existing state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 includes necessary background information about FCM, HCM and PCM. Section 3 introduces our proposed hybrid c-means classification framework. Section 4 presents the extracted feature vector. Section 5 demonstrates the competency of the proposed approach by experimental results. Section 6 concludes the paper.

2 Background information

2.1 Fuzzy c-means algorithm

The FCM is an iterative method of clustering that allows one piece of data to belong to two or more clusters [26]. An unlabelled data set $X = \{x_1, x_2, x_3, \dots, x_n\}$ represents the feature vectors of the n items. FCM sorts the data set X into c clusters. The standard FCM objective function with the Euclidian distance metric is defined as follows:

$$J_{FCM}(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m d^2(x_k, v_i) \quad (1)$$

where $d^2(x_k, v_i)$ represents the Euclidian distance between the data point x_k and the center v_i of the i -th cluster, and u_{ik} is the degree of membership of the data x_k to the k -th cluster. The parameter m controls the fuzziness of the resulting partition, with $m \geq 1$, and c is the total number of clusters. Minimization of $J_{FCM}(U, V)$ is accomplished by repeatedly adjusting the values of u_{ik} and v_i by the following equations:

$$u_{ik} = \left[\sum_{j=1}^c \left(\frac{d^2(x_k, v_i)}{d^2(x_k, v_j)} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (2)$$

$$v_i = \frac{\sum_{k=1}^n u_{ik}^m x_k}{\sum_{k=1}^n u_{ik}^m} \quad (3)$$

The iteration of the FCM algorithm is terminated when the ending condition $\max_{1 \leq i \leq c} \left\{ \text{abs} \left(v_i^{(t)} - v_i^{(t-1)} \right) \right\} < \varepsilon$ is satisfied, where $v_i^{(t-1)}$ is the center of the previous iteration, $\text{abs}(\)$ indicates the absolute value, and ε is the predefined termination threshold. Finally, all data points are distributed into clusters according to the maximum membership u_{ik} .

2.2 Hard c-means algorithm

The HCM algorithm is also an iterative method of clustering that allows one piece of data to belong to only one cluster [24]. The standard HCM objective function with the Euclidian distance metric is defined as follows:

$$J_{HCM}(H, V) = \sum_{i=1}^c \sum_{k=1}^n h_{ik} d^2(x_k, v_i) \tag{4}$$

$$h_{ik} = \begin{cases} 1 & \text{if } i = w_k \\ 0 & \text{if } i \neq w_k \end{cases} \tag{5}$$

where $w_k = \arg \min_i \{d^2(x_k, v_i) | i = 1, 2, \dots, c\}$ is the index of the cluster which wins the competition for the vector x_k . Local minimization of J_{HCM} is accomplished by repeatedly adjusting the values of h_{ik} by (5) and v_i by the mean of all data points in the cluster i .

2.3 Possibilistic c-means algorithm

The PCM algorithm proposed in [21] is stated by the standard objective function as follows:

$$J_{PCM}(T, V) = \sum_{i=1}^c \sum_{k=1}^n t_{ik}^p d^2(x_k, v_i) + \sum_{i=1}^c \eta_i \sum_{k=1}^n (1 - t_{ik})^p \tag{6}$$

where t_{ik} is the degree of typicality of the data x_k to the k -th cluster with $p \geq 1$, and η_i is a user-defined constant. The authors proposed a mathematical model to calculate η_i as follows:

$$\eta_i = K \frac{\sum_{k=1}^n t_{ik}^p \cdot d^2(x_k, v_i)}{\sum_{k=1}^n t_{ik}^p}, \quad K > 0, \tag{7}$$

where the most common choice is $K=1$.

Local minimization of the objective function $J_{PCM}(T, V)$ is accomplished by repeatedly adjusting the values of t_{ik} and v_i according to the following equations:

$$t_{ik} = \left[1 + \left(\frac{d^2(x_k, v_i)}{\eta_i} \right)^{\frac{1}{p-1}} \right]^{-1} \tag{8}$$

$$v_i = \frac{\sum_{k=1}^n t_{ik}^p \cdot x_k}{\sum_{k=1}^n t_{ik}^p}, \quad 1 \leq i \leq c. \tag{9}$$

The PCM algorithm is terminated in the same way of FCM and all data points are distributed into clusters according to the maximum typicality t_{ik} .

3 Proposed audio classification framework

3.1 The hybrid c-means approach

The FCM algorithm exhibits shortcomings regarding sensitivity to outliers and slow convergence speed, as discussed in Section I. In order to address these two problems, the PFCM presented in [22] illustrated the mixture of FCM and PCM along with two controlling parameters, and the OSFCM in [24] presented the mixture of FCM and HCM along with a controlling parameter. In order to address both problems, the core definitions of the objectives of fuzzy, hard, and possibilistic criteria together that are shown in (1), (4), and (6), respectively, were considered and the fusion of objectives was achieved by applying two controlling parameters between these three objectives. In order to derive the unified objective function of the hybrid c-means approach for audio classification, a fusion model was adopted [20], which is:

$$J_{Unified}(U, H, T, V) = abJ_{FCM} + (1-b)J_{PCM} + b(1-a)J_{HCM} \tag{10}$$

where a and b are two tradeoff parameters within the range $[0, 1]$ to control the mixture of FCM, PCM, and HCM objectives. These parameters need to be set analytically. Here, a is responsible for the FCM-HCM mixture, and b is responsible for the presence of PCM. As observed in (10), higher values of a and b indicate higher participation of the FCM objective to constitute the proposed objective by minimizing the participation of the PCM and the HCM objectives. On the other hand, lower value of a indicates higher participation of the HCM objective and lower value of b indicates higher participation of the PCM objective. The effect of these two parameters in constituting the hybrid objective function is described in the next subsection.

From (10), the objective function can be derived as follows:

$$J_{Unified} = \sum_{i=1}^c \sum_{k=1}^n [abu_{ik}^m + (1-b)t_{ik}^p + b(1-a)h_{ik}] d^2(x_k, v_i) + (1-b) \sum_{i=1}^c \eta_i \sum_{k=1}^n (1-t_{ik})^p \tag{11}$$

where, h_{ik} , u_{ik} , and t_{ik} are the degrees of the membership of input vector x_k in cluster v_i , with the usual restrictions according to HCM, FCM, and PCM criteria, respectively. The parameters m and p are the exponents of the fuzzy and possibilistic terms and restricted by $m > 1$ and $p > 1$. The parameter η_i is the variance of the possibilistic term calculated by (7).

Minimization of $J_{Unified}(U, H, T, V)$ is accomplished by repeatedly adjusting the values of u_{ik} , h_{ik} , t_{ik} , and v_i . We adjust the values of u_{ik} , h_{ik} , and t_{ik} using (2), (5), and (8), respectively. Cluster prototypes are updated according to the following equation:

$$v_i = \frac{\sum_{k=1}^n [abu_{ik}^m + (1-b)t_{ik}^p + b(1-a)h_{ik}]x_k}{\sum_{k=1}^n [abu_{ik}^m + (1-b)t_{ik}^p + b(1-a)h_{ik}]}, 1 \leq i \leq c \quad (12)$$

The iteration of the proposed hybrid c-means model is terminated when the ending condition $\max_{1 \leq i \leq c} \{abs(v_i^{(t)} - v_i^{(t-1)})\} < \epsilon$ is satisfied. Finally, all data points are distributed into clusters according to the maximum membership u_{ik} . We select the U matrix for final membership due to the proficiency of FCM over hard and possibilistic representations [20].

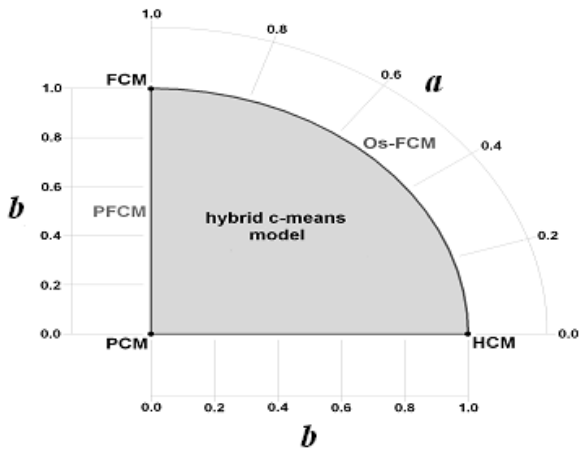


Fig 1. Analysis on the influence of controlling parameters in the hybrid objective function [20]

3.2 Analysis of the tradeoff parameters of the hybrid c-means approach

The performance of the objective function in (11) can be analyzed by the influence of fuzzy, hard, and possibilistic criteria on the hybrid objective through changing the values of the controlling parameters a and b . From the definition of the hybrid objective function in the Section 3.1, it is noted that the values of a and b are within the range $[0, 1]$. Fig. 1 illustrates the impact of a and b on the behavior of hybrid

objective and it can be observed that the value of a controls the mixture of the FCM and HCM, and the value of b controls the mixture of the FCM-PCM and HCM-PCM. Special cases occur at the boundaries and corners, such that $b=0$ makes the hybrid objective function equivalent to the PCM, $b=1$ and $a=0$ corresponds to the HCM, $b=1$ and $a=1$ is the FCM, $a=1$ and $b=[0, 1]$ is the PFCM, and $b=1$ and $a=[0, 1]$ is the OSFCM. Thus, the proposed hybrid objective mixes the objectives of the FCM, HCM, and PCM under the control of the parameters a and b . To solve the problem of sensitivity to outliers, it was necessary to ensure an appropriate mixture of FCM and PCM [22]. In addition, to solve the problem of slow convergence speed, an appropriate mixture of FCM and HCM must be ensured [24]. Fig. 1 shows that an effective mixture of FCM-PCM and FCM-HCM can be achieved by carefully selecting the values of a and b . Thus, we can overcome the problems of the conventional FCM by the hybrid c-means approach.

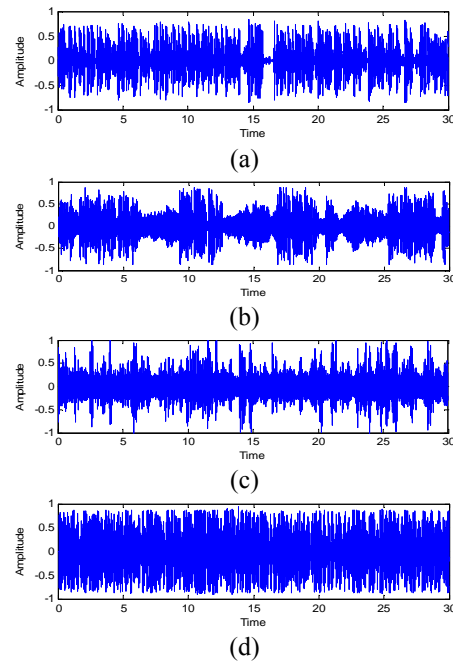


Fig 2. Sample audio signals representing different audio groups. (a) Speech, (b) speech-with-music, (c) speech-with-noise, and (d) music

3.3 Audio classification by using the hybrid c-means clustering approach

The steps of the proposed audio classification framework based on the heterogeneous c-means approach are summarized as follows:

1. Acquire audio clips and extract feature vectors as data elements.
2. Distribute the data elements of the audio clips into a data set X and initiate the center values $V^0 = (v_1^0, v_2^0, \dots, v_c^0)$.

3. Choose the values of the exponents m and p , and the tradeoff parameters a and b .
4. Calculate the values of u_{ik} , h_{ik} , η_i , and t_{ik} from (2), (5), (7), and (8), respectively.
5. Calculate the new center values of the clusters using (12).
6. Evaluate the termination condition $\max_{1 \leq i \leq c} \{abs(v_i^t - v_i^{t-1})\} < \varepsilon$. The process is finished if this condition is satisfied; otherwise repeat the process starting from step 4.
7. Assign each data element to clusters according to their maximum memberships in $U_{C \times N}$.
8. Determine the genre of each signal by resolving the corresponding cluster type through kNN.

The proposed framework is further used for audio signals classification.

4 Feature extraction

Fig. 2 illustrates sample audio signals representing different audio groups. We observe the variations in the pattern of the audio signals in different genres, which can be expressed by audio features. Individual contribution of different time, frequency and coefficient domains features in classification of broad audio genres was investigated in [11, 13, 27]. In this paper, the following 36 features which showed superior classification ability as demonstrated in the previous works were used in order to construct an efficient feature vector for the proposed classification framework.

In the time domain:

1. Mean and variances of fraction of low energy frames.
2. Mean and variances of zero crossings.

In the frequency domain:

3. Mean and variances of spectral centroid.
4. Mean and variances of spectral entropy.
5. Mean and variances of spectral roll-off.

In the coefficient domain:

6. Mean and variances of first 13 mel-frequency cepstral coefficients (MFCC).

5 Experimental result

This section evaluates the performance of the proposed approach in audio classification. First, we describe the experimental environment as well as datasets used for the performance evaluation. The comparisons of classification accuracy and computational complexity of different classifiers are also presented and discussed.

5.1 Experimental environment

The performance of the proposed approach was evaluated using a graphical user interface (GUI) simulator, and empirical values were used to set parameters such as fuzzy weighting exponents, $m=2.0$; possibilistic weighting

exponents, $p=2.0$; number of nearest neighbors in kNN, $k=30$; and termination threshold, $\varepsilon=0.001$. An extensive analysis was made within the range of values [0, 1] and the tradeoff parameters set to $a=0.84$ and $b=0.90$ in order to achieve an appropriate mixture of FCM, PCM and HCM for broad audio genre classification. A benchmark database was employed that was previously used in [3, 25, 27]. This database was created by utilizing a number of audio samples obtained from TV programs or internet, including male speech, female speech, and conversations for SP, instrumental as well as vocal songs of different music genres for MS, and signals with different levels of noises for SN audio signals. The database contains four datasets- DS1, DS2, DS3 and DS4 with each containing 150-clips (30-clips from each group). The performance of the proposed approach was compared with the following state-of-the-art approaches:

1. *Approach1*: Conventional FCM-based method [14].
2. *Approach2*: K-means based method [16].
3. *Approach3*: FCM with divergence kernel (FCMDK) method [17].
4. *Approach4*: Temporally weighted FCM method [19].

The accuracy of the proposed approach was compared with the aforementioned four approaches by the following parameter, where higher value indicates higher performance [19]:

$$\text{Precision rate, PR} = \frac{\text{Number of correctly classified audio clips}}{\text{Number of all audio clips}} \times 100\% \quad (13)$$

The percentage of the performance improvement for the proposed approach over a previous approach was calculated by the following well-known mathematical model:

$$\text{Enhancement} = \frac{\text{Present performance} - \text{Previous performance}}{\text{Previous performance}} \times 100\% \quad (14)$$

where present and previous performance requests the performances of the proposed and previously reported approaches, respectively.

In addition, the convergence speed (i.e. computational complexity) of the proposed classification framework was compared with the conventional FCM in terms of number of iterations needed to converge to the solution.

5.2 Performance evaluation

The experimental results of classification by the proposed hybrid c-means approach are summarized in Table 1, which was evaluated on all 4-datasets. Manually classified clips are by row and the results of the proposed approach are by column. One significant point can be inferred from the presented confusion matrixes, that is the likeliness of SN with SP, and SM with MS. Although the proposed approach achieved considerable success in discrimination of other groups, most of the errors occurred from these audio signals. It was observed that SN with SP type misclassification

occurred when the noise strength is very low in SN. On the other hand, SM with MS-type misclassification mostly occurred when the strength of the musical component was higher than that of the speech component in SM, or lack of continuous strings in MS clips. In addition, we assume that the selected feature vector cannot efficiently represent the characteristics of music component of very low strength. However, the discrimination accuracy within other genres was very high.

Table 1: Confusion matrix of audio genres by using the proposed approach on all datasets

Genres	SP	SN	SL	SM	MS
SP	112	3	0	3	2
SN	21	85	0	0	14
SL	0	0	120	0	0
SM	0	8	0	85	27
MS	1	7	0	23	89

The classification performance of the proposed approach was compared with the four state-of-the-art methods listed in Section 5.1. The results are summarized in Table 2, where the entries represent the number of correctly classified audio clips out of the 150 clips in each dataset. It is observed that *Approach 1* exhibited similar results to *Approach 4* due to the utilization of the same feature vectors for classification in the corresponding previous researches. *Approach 3* (a recently proposed method) showed very poor results due to the problem of coincidence clusters. Fig. 3 presents the performance comparison in terms of the parameter PR. The demonstrated performances of the proposed approach in all datasets are considerably higher than the other methods as presented in Table 2 and Fig. 3, which is due to the utilization of an efficient classification framework along with an appropriate feature vector. An overall accuracy of 81.83% has been achieved through the proposed approach by outperforming the previously reported works. In addition, the percentage of accuracy improvement by the proposed approach over the conventional FCM-based approach (*Approach 1*) was calculated by (14) and an accuracy improvement of 46.99% was achieved, which implies the competency of the proposed approach over FCM-based approach in accuracy of audio classification.

Table 2: Comparison of audio classification results of different algorithms on all datasets in terms of number of correctly classified audio clips

Approach/ Dataset	DS1	DS2	DS3	DS4
<i>Approach 1</i>	82	84	81	87
<i>Approach 2</i>	104	91	97	95
<i>Approach 3</i>	84	84	86	82
<i>Approach 4</i>	82	84	81	87
<i>The Proposed Approach</i>	127	118	127	119

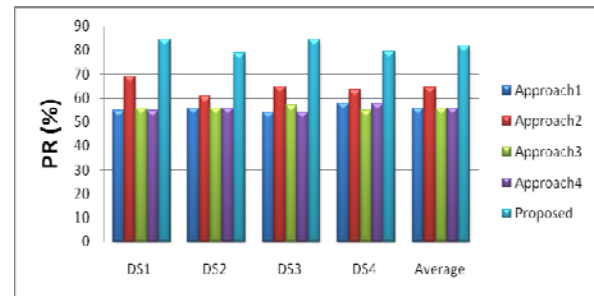


Fig 3. Classification performance comparison among different approaches in terms of PR

Table 3 presents the number of iterations required to classify signals by repeatedly minimizing the objectives of the FCM and the proposed hybrid c-means approaches. As the required number of iterations can vary depending upon the random initialization of cluster centers, thus the entries for the datasets in Table 3 were computed by running each algorithm on each dataset 5 times and then calculating the mean of number of iterations required in all runs. Additionally, as the length of the feature vector can influence the required number of iterations for convergence and a 36-elements feature vector was utilized instead of 5-elements feature vector of conventional FCM-based method, thus both classifiers were tested on both feature vectors. From the results it is observed that the proposed approach outperformed the FCM on both feature vectors in terms of required number of iterations due to the quicker convergence to the solutions. In order to calculate the percentage of time complexity reduction by (14), the approaches in first row (*Approach 1*) and fourth row (the proposed approach) from Table 3 were used. A reduction of computational complexity by 40.78% was achieved.

Table 3: Comparison of audio classification results of different algorithms on all datasets

Approach/ Dataset	DS1	DS2	DS3	DS4	Average Iterations
Conventional FCM with Nitanda's (2006) feature vector	25.80	25.80	58.40	28.40	31.75
Conventional FCM with our feature vector	42.00	46.60	43.4	53.8	41.95
The proposed approach with Nitanda's (2006) feature vector	22.00	22.20	18.40	16.00	19.65
The proposed approach with our feature vector	20.40	19.20	18.60	17.00	18.8

It is noted that the aforementioned improved accuracy of classification and reduced number of required iterations to converge to the optimal solution by the hybrid c-means approach implied the overcoming of drawbacks of the FCM through the utilization of possibilistic and hard criteria along with fuzzy criterion. It also demonstrated the competency of the proposed classification approach for broad audio genres classification.

6 Conclusions

A hybrid c-means based classification approach was proposed for broad audio genre classification by utilizing a linear combination of objective functions of FCM, PCM, and HCM along with two tradeoff parameters to control the relative strength of fuzzy, possibilistic and hard criteria of clustering in order to overcome the drawbacks of sensitivity to outliers and slow convergence speed of FCM. The utilization of the hybrid c-means concept along with an appropriate feature vector increased the robustness and accuracy in classification by overcoming the limitations of the FCM with the assurance of lower computational cost. The proposed approach was compared with some state-of-the-art audio classification approaches to validate the competency. Experimental results indicate that the proposed approach outperforms conventional FCM by exhibiting an accuracy improvement of 46.99% and a 40.78% improvement in convergence speed in broad audio genre classification. However, there still exist some performance degradations due to the lack of discrimination-ability of speech-with-music and speech-with-noise with other genres. This problem might not be related to the proposed classification framework, but might appear from the extracted features. Future analysis should focus on devising an efficient feature vector to overcome this problem of likelihood in different genres.

7 References

- [1] A. H. Mohammad and J. -M. Kim, "An enhanced fuzzy c-means algorithm for audio segmentation and classification," *Multimedia Tools and Applications*, vol. 63, no. 2, pp 485–500, 2013.
- [2] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: current directions and future challenges," *Proceedings of IEEE*, vol. 96, no.4, pp. 668–696, 2008.
- [3] A. H. Mohammad, S. Cho, and J. -M. Kim, "Primitive audio genre classification: an investigation of feature vector optimization," *Information Journal*, vol. 15, no. 5, pp. 1875–1887, 2012.
- [4] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp 27–36, 1996.
- [5] G. Tzanetakis and P. Cook, "Music genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp 293–302, 2002.
- [6] J. Saunders, "Real-time discrimination of broadcast speech/music," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Atlanta, USA, May 1996, pp. 993–996.
- [7] L. Lu, H. J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.
- [8] L. Lu, S. Z. Li, and H. J. Zhang, "Context-based audio segmentation using support vector machines," *IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, 2001, pp. 191–194, 2001.
- [9] L. Lu, H. J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.
- [10] Y. Zhu, Z. Ming, and Q. Huang, "SVM-based audio classification for content-based multimedia retrieval," *Lecture Notes on Computer Science*, vol. 4577, pp. 474–482, 2007.
- [11] P. Krishnamoorthy and S. Kumar, "Hierarchical audio content classification system using an optimal feature selection algorithm," *Multimedia Tools and Applications*, vol. 54, no. 2, pp. 415–444, 2011.
- [12] M. Liu, C. Wan, and L. Wang, "Content-based audio classification and retrieval using a fuzzy logic system: towards multimedia search engines," *Soft Computing*, vol. 6, no. 5, pp. 357–364, 2002.
- [13] M. K. S. Khan and W. G. A. Khatib, "Machine-learning based classification of speech and music," *Multimedia Systems*, vol. 12, no. 1, pp. 55–67, 2006.
- [14] N. Nitanda, M. Haseyama, and H. Kitajima, "Audio signal segmentation and classification using fuzzy c-means clustering," *Systems and Computers in Japan*, vol. 37, no. 4, pp. 23–34, 2006.
- [15] D. C. Park, D. H. Nguyen, S. H. Beack, and S. Park, "Classification of audio signals using gradient-based fuzzy c-means algorithm with divergence measure," *Lecture Notes on Computer Science*, vol. 3767, pp. 698–708, 2005.
- [16] D. C. Park, C. N. Tran, B. J. Min, and S. Park, "Modeling and classification of audio signals using gradient-based fuzzy c-means algorithm with a mercer kernel," *Lecture Notes on Computer Science*, vol. 4099, pp. 1104–1108, 2006.

- [17] D. C. Park, "Classification of audio signals using fuzzy c-means with divergence-based kernel," *Pattern Recognition Letters*, vol. 30, no. 9, pp. 794–798, 2009.
- [18] H. V. Luong, C. –H. Kim, and J. –M. Kim, "Classification of audio signals using generalized spatial fuzzy clustering," *Journal of Acoustical Society of America*, vol. 125, no. 4, pp. 2699, 2009.
- [19] N. T. T. Nguyen, A. H. Mohammad, C. –H. Kim, J. –M. Kim, "Audio segmentation and classification using a temporally weighted fuzzy c-means algorithm," *Lecture Notes on Computer Science*, vol. 6676, pp. 447–456, 2011.
- [20] L. Szilagyi, S. M. Szilagyi, and Z. Benyo, "A unified approach to c-means clustering models," *IEEE International Conference on Fuzzy Systems*, Jeju, Korea, Aug. 2009, pp. 456–461.
- [21] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.
- [22] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, 2005.
- [23] J. L. Fan, W. Z. Zhen, and W. X. Xie, "Suppressed fuzzy c-means clustering algorithm," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1607–1612, 2003.
- [24] L. Szilagyi, S. M. Szilagyi, and Z. Benyo, "Analytical and numerical evaluation of the suppressed fuzzy c-means algorithm," *Lecture Notes on Computer Science*, vol. 5285, pp. 146–157, 2008.
- [25] A. H. Mohammad, S. Cho, J. –M. Kim, "An analysis of feature space for audio classification and retrieval using a fuzzy c-means algorithm," *International Conference on Computing and Convergence Technology*, Seoul, Korea, Oct. 2011, pp. 427–436.
- [26] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal, "Fuzzy models and algorithms for pattern recognition and image processing." Springer, 2005.
- [27] A. H. Mohammad and J. –M. Kim, "An analysis of content-based classification of audio signals using a fuzzy c-means algorithm," *Multimedia Tools and Applications*, vol. 63, pp. 77–92, 2013.

Surface registration by markers guided non-rigid Iterative Closest Points algorithm

Dominik Spinczyk

Faculty of Biomedical Engineering, Silesian University of Technology, Gliwice, Silesia, Poland

Abstract - The problem of matching irregular surfaces was tested with additional markers as landmarks for the extension of the non-rigid Iterative Closest Points (ICP) algorithm. The general idea of presented approach was to take into account knowledge about markers' positions not only in computing transformation phase but also in finding correspondence phase in every algorithm's iteration. Four variants of retrieving correspondence were implemented and compared: the Euclidean distance, normal vectors with initial rigid registration, static and dynamic markers vectors. To evaluate different manner of computing correspondence the average correspondence assignment error of points nearest to the markers and the number of correspondences for every target points were defined. The presented approach was evaluated using abdominal surfaces data set, consist of captured clouds of points during free breathing of 6 volunteers. The modifications significantly improved results. To make the proposed changes more universal k-nearest neighbor method and radius constraint could be used.

Keywords: markers guided geometry-based registration, finding correspondences, non-rigid Iterative Closest Points, surface registration quality, surface registration.

1 Introduction

Registration is a process to find correspondence between data sets, and generally could be divided into geometry-based or intensity-based methods [1]. Nowadays, in many multimedia application input data set are represented as point cloud (segmented surfaces of the objects etc.). Then in the processing pipeline data sets should be register, to find the correspondence between data sets. The most popular approach, in this case, is Iterative Closest Point algorithm (ICP) [2]. This algorithm was proposed by a few researchers independently Besl [3] and Chen [4]. The ICP is an iterative algorithm and consists of two steps. The first step is to find a correspondence between target and source points, based on Euclidean distance between points. In the second step, the updated version of result transformation is calculated using equation:

$$f(S, T) = \frac{1}{N_s} \sum_{i=1}^{N_s} \|T_i - Rot(S_i) - Trans(S_i)\|^2 \quad (1)$$

where: T, S are target and source set of points,

N_s is number of source points (equals number of target points), and Rot, Trans are rotation and translation components of final transformation.

The updated version of final transformation in current iteration is based on close-form solution of mean square error problem. The classical approach used only one rigid or affine transformation for whole data sets. In literature the description of disadvantages of the classical ICP approach Rusinkiewicz [5] can be found:

- problem of finding global minimum of cost function depends on an initial guess of final transformation,
- the algorithm is sensitive to improper correspondences,
- long time of computation - one of the most time-consuming operation is retrieving correspondences.

Due to these disadvantages researchers proposed a lot of classical approach rectifications:

- registration only subsets of points,
- improvement of finding correspondence problem,
- quantity measure of proper correspondence,
- elimination of improper correspondence,
- modification of computing minimum of cost function.

The standard ICP approach cannot be used to track surface of objects that change their shape in time. Amberg [6] proposed non-rigid version of ICP, by the following equation:

$$E(X) = E_d(X) + \alpha E_s(X) + \beta E_l(X) \quad (2)$$

where:

X is the unknowns are organised in a $4n \times 3$ transformation matrix,

$E_d(X)$ is distance measure between all targets points and transformed source points, in contrast to classical ICP. X is not a single rotation or translation but a collection of affine transformation for each point,

$E_s(X)$ is stiffness regularization, topology matrix is created based on points neighbourhood to preserve the shape of object during iterations; we use square matrix topology (every point has four neighbours). α is stiffness vector, which influences the flexibility of cloud shape,

$\beta E_s(X)$ is a factor used for guiding registration, based on known position of landmarks in source and target sets of points. β is an weighting factor, used to fade out the importance of landmarks towards the end of the registration process.

The implemented non-rigid ICP algorithm consists of two iterative loops. In the outer loop, the stiffness factor α is gradually decreased with uniform steps, starting from higher values, which enables recovery of an initial rigid global alignment, to lower values, allowing for more localized deformations. For a given value of α , the problem is solved iteratively in the inner loop. The condition of changing stiffness vector is threshold norm of transformation difference from adjoining iterations. In our implementation β is constant and equals one. The above equation can be transformed into the system of linear equations, which is solved by computing pseudo-inverse matrix (see [6] for details). This is the iterative algorithm, where each iteration consists of two main steps, namely finding correspondences between source and target points and computing affine transformations for each source point. If the second step is modified by the solution proposed by Amberg, that causes better results, corresponding problem remains critical for final results.

2 Material and Methods

Improvement of finding correspondences was implemented. Classically finding correspondences is done by searching Euclidean distance between closest points in source and target or in normal vector of source point direction. The general idea of presented approach was to take into account knowledge about markers' positions not only in computing transformation phase but also in finding correspondence phase in every algorithm's iteration. Decision to test a few approaches of finding correspondences was done:

- searching along normal vectors in source points, following the initial rigid registration based on Horn algorithm [7],
- along static marker vector displacement, where marker vectors are calculated only once at the preliminary stage. Marker vector is defined by positions of specific marker in source and target point cloud,

- along dynamic marker vector displacement, where marker vectors are calculated in each iteration based on constant positions of nearest marker points in matrix topology. Transformed source point in each iteration is treated as a new origin of dynamic marker vector.

Classical Euclidean distance is treated as baseline to compare the obtained results. Generally it is challenging to verify registration approach. We used global and local criteria for evaluation:

- global measurement: average distances between nearest source and target points, average distances between correspondences,
- local measurement – quality of correspondences: average correspondence assignment error of points nearest to the markers and the number of correspondences for every target points.

Data set consists of abdomen surfaces acquired by 6 volunteers on free breathing using Time-of-Flight sensor Mesa SR4000 [8]. Intensity map example of input data is presented in Fig. 1. As markers 15mm white squares attached to the abdomen were used, which corners were manually segmented by two users.



Figure 1. Example of input data: intensity map for ToF camera of abdomen with nine square markers.

3 Results

For the different methods of finding correspondences, evaluation scores: surface distances, correspondence distances and average marker error, are presented in tables 1 and 2.

Table 1. Surface distances for four variants of ICP: the Euclidean distance (E), normal vectors with initial rigid registration (NH), static (SM) and dynamic markers vectors (DM).

ID	Surface Distance [mm]				
	Initial	E	NH	SM	DM
1	5.83	0,21	0.6	0.69	0.63
2	12.04	0.12	0.74	0.87	0.61
3	25.16	0.03	0.03	0.06	3.43
4	19.84	0.14	1.04	0.51	3.56
5	5.21	0.004	0.21	0.28	0.28
6	10.76	0.16	0.15	0.09	1.94

4 Discussion and Conclusions

The implemented non-rigid ICP algorithm showed average residual distance 0.68mm (Euclidean distance not included). A further analysis of registration accuracy was focused on finding correspondence problem. Four methods for this problem were tested: Euclidean distance treated as base line, normal shooting with initial rigid registration – marker based Horn algorithm, static marker vectors (computing only one at the beginning of registration process) and dynamic marker vectors (computing in every iteration). For “near” cloud, where stiffness vector is constant for almost every iterations in non-rigid ICP, Euclidean distance are good enough. Unlike “near” clouds, “far” clouds, where stiffness vectors are changing for few iterations, Euclidean distance seems to be not enough. There are a lot of gaps in registered source cloud – Fig. 2. To improve it, static and dynamic marker vectors were proposed. If marker is not only used in computing transformation step but also in computing correspondences step for each iteration, correspondence assignment error of points nearest to the markers decreased from 5.4 to 2.0 of confused neighbors. Normal shooting approach was also evaluated, but results were worse results than other cases, while combination normal shooting and initial rigid registration significantly improved results – Fig. 2. To use Horn algorithm at least three non collinear corresponding points in source and target should be known. It helps allows to overcome the problem of the relative displacement of the source and target point clouds, which is not taking into account when Euclidean distance is used.

Because it is difficult to measure directly the quality of correspondences, observation was proposed in a few steps. Correspondence map (Fig. 2) showed spatial distribution of the feature, number of correspondences assigned to every

target point (desirable value is 1). It is easier to compare correspondence map globally with different cases using correspondence map histogram (Fig. 3). Average correspondence assignment error of points nearest to the markers allows to measure the quality of correspondence points from cloud, which are nearest to the markers. To make the proposed changes of finding correspondences more universal k-nearest neighbor method and radius constraint could be used, to apply marker information not to every point in cloud but only to for the nearest points to the markers. For points which are not the nearest to the marker, Euclidean distance or normal shooting could be used.

Presented approach may be used in different medical, entertainment and industrial applications, where non-rigid point clouds should be registered, when initial relative position of clouds is that finding correspondences by Euclidean distance or normal shooting is not enough. The proposed changes do not introduce complex calculations. Initial calculation of rigid registration allows to solve the problem of unknown transformation matrix initialization. Comparing to classical non-rigid ICP the disadvantage of proposed approach is that initial corresponding positions of markers in source and target point clouds are needed.

5 References

- [1] M. Wyawahare, Dr. P. Patil, and H. Abhyankar: Image Registration Techniques: An overview. International Journal of Signal Processing, Image Processing and Pattern Recognition 2009;2(3):11-28.
- [2] J. Salvia, C. Mataboscha, D. Fofib and J. Forest: A review of recent range image registration methods with accuracy evaluation. Image and Vision Computing 2007;25:578–596.
- [3] P. Besl, and N. McKay: A method for registration of 3D shapes. Pattern Analysis Machine Intelligence 1992;14(2):239-56.
- [4] Y. Chen and G. Medioni: Object modeling by registration of multiple range images Image Vision Comput 1992;10(3):145-55.
- [5] Sz. Rusinkiewicz and M. Levoy: Efficeint Variants of the ICP Algorithm. Proceedings of the of 3rd Int. Conf. on 3-D Digital Imaging and Modeling. Stanford Univ., CA, USA 2001:145–52.
- [6] B. Amberg, S. Romdhani and T. Vetter: Optimal Step Nonrigid ICP Algorithms for Surface Registration. Proceedings of IEEE Conference of Computer Vision and Pattern Recognition CVPR 2007:1-8.
- [7] B. Horn, H.Hilden and S. Negahdaripour: Closed form solution of absolute orientation using orthonormal matrices. J Opt Soc Am A. 1988;5:1127-35.
- [8] Mesa-Imaging - manufactor website: SR4000 Data Sheet <http://www.mesa-imaging.ch/swissranger4000.php>.

Acknowledgment

The study was supported by National Science Center, Polad, Grant No UMO-2-12/05/B/ST7/02136.

Table 2. Correspondence distances and average marker error for four variants of ICP: the Euclidean distance (E), normal vectors with initial rigid registration (NH), static (SM) and dynamic markers vectors (DM).

ID	Correspondence Distance [mm]				Marker Error [number of unit]			
	E	NH	SM	DM	E	NH	SM	DM
1	0.26	0.18	0.25	0.58	12.42	10.81	4.45	0.76
2	0.12	0.27	0.33	0.87	7.02	6.28	2.12	1.49
3	0.04	0.03	0.03	1.7	2.57	2.9	1.86	0.44
4	0.14	0.22	0.15	0.88	4.3	4.63	3.6	0.72
5	0.04	0.07	0.1	0.1	4.05	2.61	3.11	4.69
6	0.16	0.07	0.06	0.53	2.25	1.89	0.75	0.09

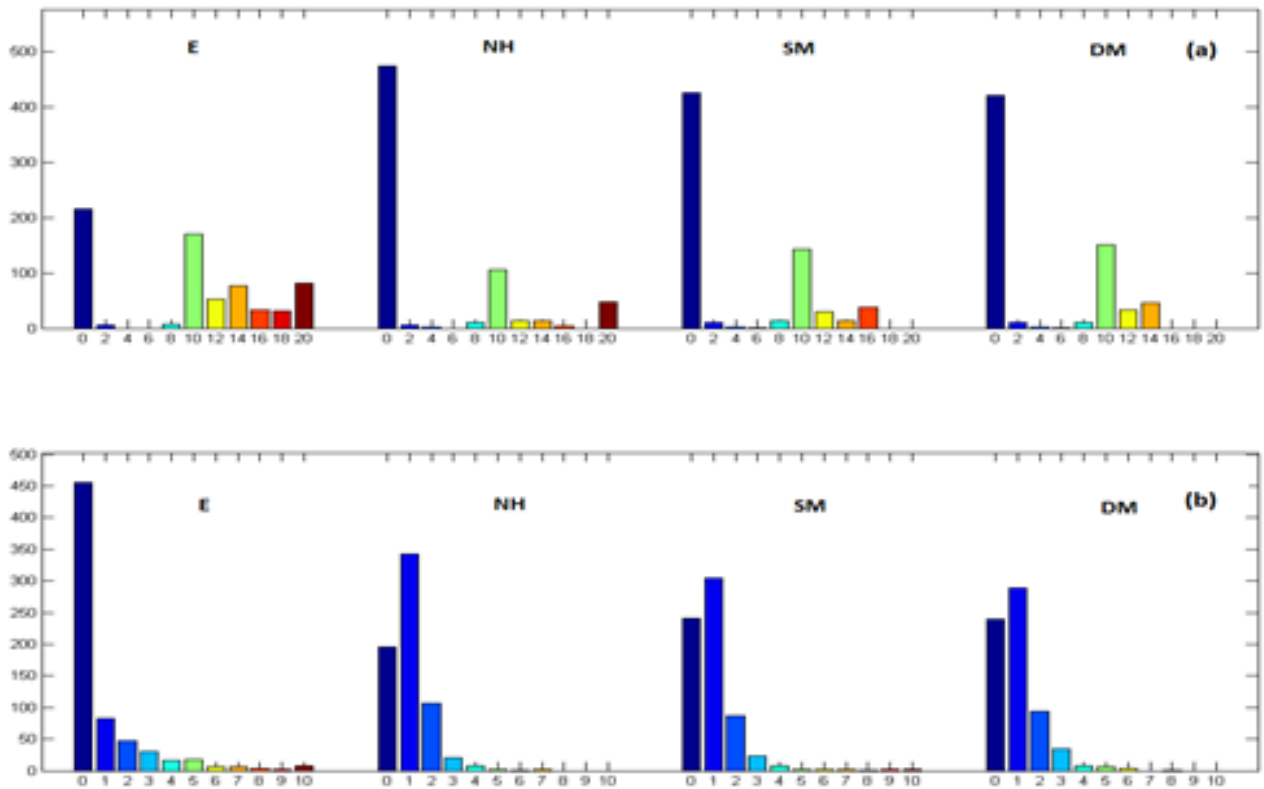


Figure 3. Distance map histogram [mm] (a) and correspondence map histogram [number of units] (b) in different modifications of ICP computing correspondence: Euclidean distance (E), normal shooting with initial rigid registration (NH), static marker vectors (SM), dynamic marker vectors (DM).

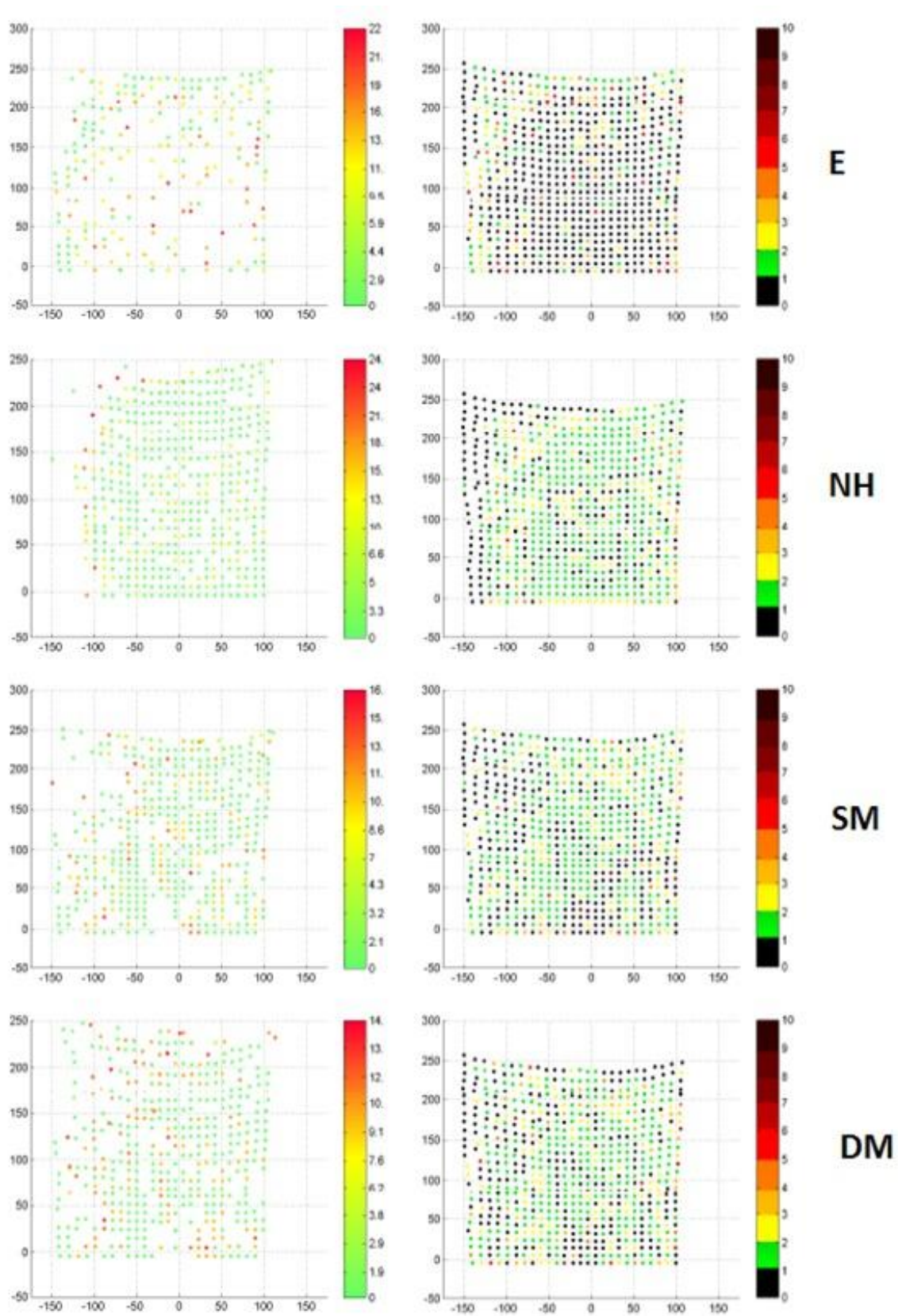


Figure 2.Distance map [mm] (left column) and correspondence map [number of units] (right column) for different modifications of ICP computing correspondence: Euclidean distance (E), normal shooting with initial rigid registration (NH), static marker vectors (SM), dynamic marker vectors (DM).

AN AFFINE SHAPE CONSTRAINT FOR GEOMETRIC ACTIVE CONTOURS

Mohamed Amine Mezghich, Mawaheb Saidani, Slim M'Hiri and Faouzi Ghorbel

GRIFT Research Group, CRISTAL Laboratory
 École Nationale des Sciences de l'Informatique, ENSI
 Campus Universitaire de la Manouba, 2010 Manouba, Tunisia
 ma.mezghich@crystal.rnu.tn, {mawaheb.saidani,slim.mhiri,faouzi.ghorbel}@ensi.rnu.tn

ABSTRACT

We intend to present in this research a new method to incorporate affine invariant geometric shape prior into a level set based active contours for robust segmentation of partially occluded object. The proposed shape constraint is defined after an affine shape alignment of the active contour and a reference shape. In order to generalize our work to the multi-references case, we present a set of complete and stable invariant shape descriptors computed using Fourier transform of the contours to choose the most suitable reference according to the evolving contour. Experiments on synthetic and real data show the ability of the proposed approach to constrain an evolving curve towards a target shapes that may be occluded and cluttered under geometric transformations.

Index Terms— Affine shape alignment, affine invariant shape descriptors, shape constraint, active contours.

1. INTRODUCTION

Precise object of interest detection is of high need in many applications of image processing such as medical imaging, tracking of moving object and 3D object reconstruction. There have been several works in this field of image segmentation and object detection. Among them we can distinguish the classical active contours which are intensity-based models [1, 2, 3, 4, 5] and the constrained models by a prior knowledge that are driven both by global geometrical [6, 7, 8, 9] or statistical [10, 11, 12, 13, 14] shape information and a local image information like gradient or curvature. There have been promising results given that the classical snake models can detect object with smooth contours whereas the level set based models have the additional ability to detect simultaneously many objects in the image. However, several work of the prior knowledge driven models reached good segmentation results by detecting partially occluded shapes presenting missing part in low contrast or very noisy image. To define prior knowledge, the authors use either shape alignment [7] or registration [9] between the contour of the target shape and the shape of reference or a distance between invariant shape descriptors like [8]. All these work manage

the case of Euclidean transformations (translation, rotation and scale factor). But in real situations, the object of interest can submit large transformations (affine or projective cases) and distortions. Hence the class of rigid transformations are not appropriate to model this type of movement between the shape of reference and the target one. To our knowledge, there has been only the work of Foulonneau et al. [15] that treats the case of affine transformations. Although this approach, which is based on the distance between an invariant set of Legendre descriptors of the target and reference shapes, presented good results, this method suffers from the instability of the Legendre moments and the order that has to be fixed empirically. Besides, the method requires a heavy execution time to achieve satisfactory results. In order to extend the work of shape priors presented in [16, 17]. We propose in this research a geometric approach to manage the class of affine transformations that can occur between the shape of interest and the reference one. We use an explicit affine shape alignment method which is based on the Fourier transform of the curve to define prior knowledge. At the beginning, we consider that the shape of reference is given, then we were based on a set of an affine invariant shape descriptors which are complete and stable to select the best template in case of many available references. The remainder of this paper is organized as follows : In Section 2, we will recall the used shape alignment method based on Fourier transform. Then, the proposed shape priors will be presented in Section 3. Experiments will be presented and commented in Section 4. Finally, we conclude the work and highlight some perspectives in Section 5.

2. AFFINE SHAPE AND MOTION DESCRIPTION

In order to define affine invariant shape prior from a reference shape, we have to perform shape matching between the target and the template shapes F_1 and F_2 . To correctly estimate the affine motion parameters, we have to deal with the same number of points which is not relevant if we consider two different points of view. For this purpose in [18], the author proposes to use the affine reparametrization of curve which is invariant

under affine transformation. We will start by presenting the affine reparametrization procedure of a given curve. Then we will describe our method of affine motion's parameters estimation. By the last paragraph of this section, we present the set of invariant shape descriptors used to choose the best reference.

2.1. Reparametrization of closed curves

If we take the same object in two different camera sides, we found a different number of contour points in each image. Consequently, we proceed by normalization of curves under affine transformation. Given the shape front, its edge pixels are extracted and traversed to yield a discrete closed curve which is a parametric equation $\gamma(t) = (u(t), v(t))$ where $t \in \{0, \dots, N-1\}$ and $\gamma(N) = \gamma(0)$. We use the affine arc length reparametrization to normalize closed curve under affine transformation.

$$s_a(t) = \frac{1}{L_a} \int \|\gamma'(t) \wedge \gamma''(t)\|^{\frac{2}{3}} dt, \quad t \in [a, b] \quad (1)$$

$$L_a = \int \sqrt[3]{|\gamma'(t) \wedge \gamma''(t)|} dt \quad (2)$$

Where L_a is the total affine arc length of the considered curve, \wedge represents the cross product between two vectors and $\|\cdot\|$ denotes the Euclidean norm.

2.2. Contours alignment using geometrical affine parameters estimation

We consider two closed curves O_1 and O_2 which define the same shape. O_1 and O_2 are said to be related by an affine transformation if and only if

$$h(l) = \alpha A f(l + l_0) + B \quad (3)$$

where B is a translation vector, A is a linear transformation, l_0 is the shift value, α is the scale factor, f and h are the affine reparametrization of two contours having the same affine shape. The Fourier transform in this case corresponds to the Fourier coefficients of an affine arc length parametrization of a given curve. So, In Fourier space we get:

$$U_k(h) = \alpha e^{2i\pi k l_0} A U_k(f) + b \delta_k \quad (4)$$

where $U_k(h)$ and $U_k(f)$ are respectively the Fourier coefficients of f and h . So estimation of affine motion can be resumed to the estimation of its three parameters : the affine matrix A , the shift value l_0 and finally the scale factor α if we consider a normalization under translation.

2.2.1. Estimation of the scale factor α

The scale factor can be estimated using the following formula

$$\alpha^2 = \frac{\det(U_h(k), U_h^*(k))}{\det(U_f(k), U_f^*(k))} \quad (5)$$

where $(U_h(k), U_h^*(k))$ and $(U_f(k), U_f^*(k))$ are 2×2 matrix formed by Fourier coefficients on some fixed index k and U^* is the U complex conjugate.

2.2.2. Computation of the Shift value l_0

Let's consider $M_1 = (U_h(k_1), U_h(k_2))$ and $M_2 = (U_f(k_1), U_f(k_2))$

$$l_0 = \frac{\arg(\det(M_2)) - \arg(\det(M_1))}{(k_1 + k_2)} \quad (6)$$

where k_1 and k_2 are two fixed index and $\arg(U)$ is the complex argument of U .

2.2.3. Computation of the Matrix A 's parameters

Since scale factor α and shift value l_0 have been previously estimated (eq. 5) and (eq. 6), we are going to estimate affine matrix A parameters. Let O_1, O_2 be two curves and U_f, U_h respectively their Fourier coefficients. In Fourier space we have (eq. 7)

$$U_h(k) = \alpha e^{jk l_0} A U_f(k) \quad (7)$$

Using matrix, we have the following formula (eq. 8)

$$\begin{pmatrix} u_2(k) \\ v_2(k) \end{pmatrix} = \alpha e^{jk l_0} \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} u_1(k) \\ v_1(k) \end{pmatrix} \quad (8)$$

So we have to estimate a_1, a_2, a_3 and a_4 . This system can be represented as following

$$\left\{ \begin{array}{l} u_2(k_1) = \alpha e^{jk_1 l_0} a_1 u_1(k_1) + \alpha e^{jk_1 l_0} a_2 v_1(k_1) \\ v_2(k_1) = \alpha e^{jk_1 l_0} a_3 u_1(k_1) + \alpha e^{jk_1 l_0} a_4 v_1(k_1) \\ \vdots \\ u_2(k_n) = \alpha e^{jk_n l_0} a_1 u_1(k_n) + \alpha e^{jk_n l_0} a_2 v_1(k_n) \\ v_2(k_n) = \alpha e^{jk_n l_0} a_3 u_1(k_n) + \alpha e^{jk_n l_0} a_4 v_1(k_n) \end{array} \right\} \quad (9)$$

Then to estimate matrix A 's parameters (eq. 7), we have to resolve a system with $2N$ equations and 4 unknown parameters that can be written as

$$K_{2n \times 4} A_4 = U_{2n} \quad (10)$$

The solution of the system (eq. 9) can be obtained by

$$KA - U = e \quad (11)$$

so we have to minimize the quadratic error e by using the pseudo-inverse of K . We show by the following figure an example of affine motion estimation between two synthetic curves after contours re-sampling. For more affine contours matching results using the method explained above, the reader can be referred to [19, 20, 21, 22]. The final result of curves alignment after affine motion estimation are presented by (Figure.2). The estimated motion parameters are presented by Table 1.

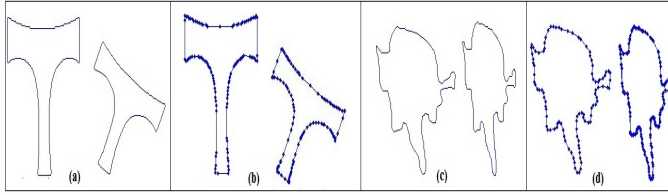


Fig. 1. Two different shapes after affine deformations ((a) and (c)) and curve resampling ((b) and (d)).

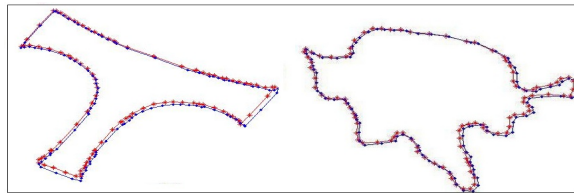


Fig. 2. Results of affine shape alignment.

	l_0	α	A matrix
Shapes of (b)	0.002	1.03	$\begin{pmatrix} 0.860 & -0.47 \\ 0.52 & 0.865 \end{pmatrix}$
Shapes of (d)	0	2.004	$\begin{pmatrix} 0.968 & 0.002 \\ 0.1037 & 0.966 \end{pmatrix}$

Table 1. The estimated affine motion parameters.

Having the parameters of the affine transformation between the two contours, we perform the alignment of the two curves to determine the regions of variability between shapes by computing the product function of the signed distance functions associated to level set functions of respectively the evolving and the reference object after alignment given by

$$f_{prod}(x, y) = f_{\phi_{ref}}(x, y) \cdot f_{\phi}(x, y), \quad (12)$$

where $f_{\phi_{ref}}$ and f_{ϕ} are the two binary images associated respectively to ϕ_{ref} and ϕ . See [16] for more details. By construction, the product function f_{prod} is negative in the areas of variability between the two binary images $f_{\phi_{ref}}$ and f_{ϕ} due to occlusion, clutter or missing parts, whereas in positive regions, the objects are similar. Thus, in what follows, we propose to update the level set function ϕ only in regions of variability between shapes to make the evolving contours overpass the spurious edges and recover the desired shapes of objects. This property recalls the Narrow Band technique used to accelerate the evolution of the level set functions [4].

2.3. Global matching using Affine Invariants Descriptors (AID)

In presence of many templates, we have to choose the most suitable one according to the evolving curve. Let α and β be positive real numbers, and k_0, k_1, k_2 and k_3 four positives

integers. Let C_n^x and C_n^y be the complex Fourier coefficients of the coordinates (u, v) , Δ denotes the determinant.

$$\Delta_n^m = \Delta \begin{vmatrix} C_n^x & C_m^x \\ C_n^y & C_m^y \end{vmatrix} \quad (13)$$

In [23], the author introduced two sets of invariant descriptors I and J which are respectively given respectively by (eq. 14) and (eq. 15).

$$I_{k_1} = |\Delta_{k_1, k_0}|, I_{k_2} = |\Delta_{k_2, k_0}|, I_k = \frac{\Delta_{k, k_0}^{k_1-k_2} \Delta_{k_1, k_0}^{k_2-k} \Delta_{k_2, k_0}^{k-k_1}}{|\Delta_{k_1, k_0}^{k_2-k-\alpha}| |\Delta_{k_2, k_0}^{k-k_1-\beta}|} \quad (14)$$

for all $k \in N^* - \{k_0, k_1, k_2\}$

$$J_{k_1} = |\Delta_{k_1, k_3}|, J_{k_2} = |\Delta_{k_2, k_3}|, J_k = \frac{\Delta_{k, k_3}^{k_1-k_2} \Delta_{k_1, k_3}^{k_2-k} \Delta_{k_2, k_3}^{k-k_1}}{|\Delta_{k_1, k_3}^{k_2-k-\alpha}| |\Delta_{k_2, k_3}^{k-k_1-\beta}|} \quad (15)$$

for all $k \in N^* - \{k_1, k_2, k_3\}$

In [19],[20], authors have shown experimentally that such descriptors are complete and stable. The completeness guarantee the uniqueness of matching, the stability gives a robustness under non linear shape distortions and numerical errors. In [23], the author demonstrates that the shape space S can be considered as a metric space with a set of metrics. Hence, the Euclidean distance (eq.16) between the set of the presented invariants can be used to compare the evolving curve and the available templates.

$$d_{\alpha}(F, H) = \|I_n(f) - I_n(h)\|_{l^{\alpha}} = \left(\sum |I_n(f) - I_n(h)|^{\alpha} \right)^{\frac{1}{\alpha}} \quad (16)$$

for any real number α such that $\alpha > 1$. Where f and h are two normalized affine arc length parametrization of two objects having respectively the shapes F and H . The shape having the minimum distance according to the evolving active contour is used as template.

3. SHAPE PRIORS FOR GEOMETRIC ACTIVE CONTOURS

Geometric active contours are iterative segmentation methods which use the Level Set approach [2] to determine the evolving front at each iteration. Several models have been proposed in literature that we can classify into edge-based or region-based active contours. In [4], the level set approach is used to model the shape of objects using an evolving front. The evolution's equation of the level set function ϕ is

$$\phi_t + F|\nabla\phi| = 0, \quad (17)$$

F is a speed function of the form $F = F_0 + F_1(K)$ where F_0 is a constant advection term equals to (± 1) depends of the object inside or outside the initial contour. The second term

is of the form $-\epsilon K$ where K is the curvature at any point and $\epsilon > 0$. To detect the objects in the image, the authors proposed to use the following function which stops the level set function's evolution at the object boundaries

$$g(x, y) = \frac{1}{1 + |\nabla G_\sigma * f(x, y)|}, \quad (18)$$

where f is the image and G_σ is a Gaussian filter with a deviation equals to σ . This stopping function has values that are closer to zero in regions of high image gradient and values that are closer to unity in regions with relatively constant intensity. Hence, the discrete evolution equation is

$$\frac{\phi^{n+1}(i, j) - \phi^n(i, j)}{\Delta t} = -g(i, j) F(i, j) |\nabla \phi^n(i, j)|, \quad (19)$$

It's obvious that the evolution is based on the stopping function g which depends on the image gradient. That's why this model leads to unsatisfactory results in presence of occlusions, low contrast and even noise. To make the level set function evolves in the regions of variability between the shape of reference and the target shape, we propose the new stopping function as follow

$$g_{shape}(x, y) = \begin{cases} 0, & \text{if } \phi_{prod}(x, y) \geq 0, \\ \text{sign}(\phi_{ref}(x, y)), & \text{else,} \end{cases} \quad (20)$$

where $\phi_{prod}(x, y) = \phi(x, y) \cdot \phi_{ref}(x, y)$, ϕ is the level set function associated to the evolving contour, while ϕ_{ref} is the level set function associated to the shape of reference after alignment. As it can be seen, the new proposed stopping function only allows for updating the level set function in the regions of variability between shapes. In these regions g_{shape} is either 1 or -1 because in the case of partial occlusions, the function is equals to 1 in order to push the edge inward (deflate) and in case of missing parts, this function is equals to -1 to push the contour towards the outside (inflate). This property recalls the Balloon snake's model proposed by Cohen in [24] in which the direction of evolution (inflate or deflate) should be precised from the beginning. In our work, the direction of evolution is huddled automatically based on the sign of ϕ_{ref} . The total discrete evolution's equation that we propose is as follows

$$\frac{\phi^{n+1}(i, j) - \phi^n(i, j)}{\Delta t} = -(w g(i, j) + (1 - w) g_{shape}(i, j)) F(i, j) |\nabla \phi^n(i, j)|, \quad (21)$$

w is a weighting factor between the image-based force and knowledge-driven force. See [16] for our proposed shape prior for a region-based active contours.

4. EXPERIMENTAL RESULTS

4.1. Robustness of the proposed shape priors

We present in Fig.3 an example of successive evolutions between several shapes of different topologies under the proposed shape priors only ($w = 0$). This simulation shows that

the proposed shape priors can well constrain an active contour to take a given shape (known as reference) and handling non trivial geometric shapes with holes and complex topologies.

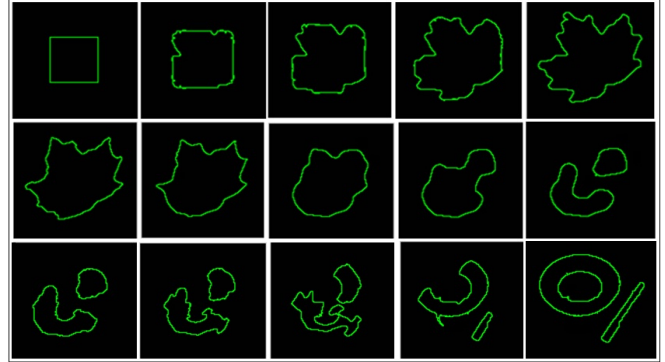


Fig. 3. Curve evolution under the proposed shape prior only.

4.2. Application to object detection

We will devote this section to present some results of object detection obtained by the proposed model in case of partially occluded object under affine transformation. We first evolve the active contour without shape prior until convergence (i.e. $w = 1$) to reduce the computational complexity and to have a good estimation of the parameters of the rigid transformation as in [11, 8]. This first result provides an initialization for the model with prior knowledge. More weight is assigned to prior knowledge (generally $w \leq 0.5$) to promote convergence toward the target shape.

Consider for this first experiment the spider object, image (a), obtained from MCD data base ¹. The (b) image is the object of interest which is obtained from (a) after affine transformation and partial occlusion. As a first step, we perform object (b) segmentation without prior knowledge. The obtained contour is then aligned with the contour of the (a) object to determine the regions of occlusions. We present in (c) the result of shape alignment. The red sampled contour corresponds to the (a) shape after the affine transformation estimation with (b) and the blue one corresponds to the (b) shape. The estimated values are $\alpha = 2, l_0 = 0, A = [0.5, 0.2, 0, 2]$. By Fig.5, we present the obtained results without and with shape prior.

We considered in this experiment four templates (a spider, a chopper, a device and a bird), see Fig.6. We were based on the set of the presented invariant descriptors to choose the suitable reference according to the occluded spider (Image (b) of Fig.6).

Table 2 presents the obtained Euclidean distance between the target shape and the available templates. We notice that the minimum distance can be easily identified because the suitable form in this experiment is distinguishable. In the

¹<http://vision.ece.ucsb.edu/~zuliani/Research/MCD/MCD.shtml>

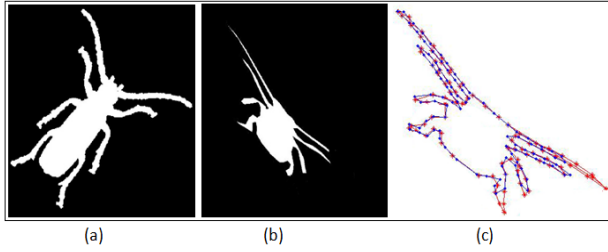


Fig. 4. Spider's shapes alignment.

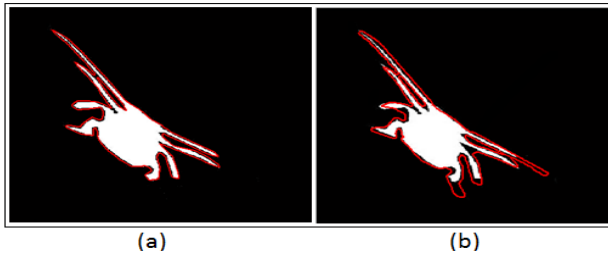


Fig. 5. Detection without, image (a), and with, image (b), affine prior knowledge.

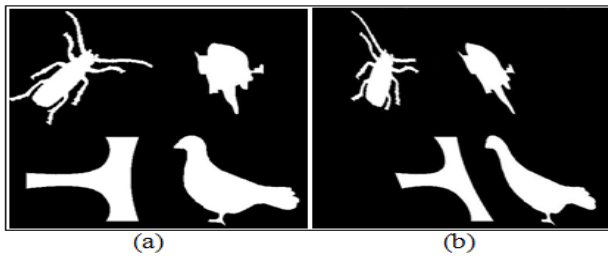


Fig. 6. Image (a): The available templates, Image (b): the transformed object.

	Spider	Chopper	Device	Bird
The occluded spider	0.037	0.46	0.44	0.53

Table 2. Distances between the occluded spider and the used templates.

next experiment, a mosaic application is considered. The mosaic images that are considered are taken from the Bardo Museum of Tunisia which contains the biggest collection of mosaic images in the world. In mosaic images, objects are composed of tessellas and are often partially occluded as it is shown by images (c) and (d), Fig.7. So given that in mosaic images the object are often repeated and in order to study the robustness of our method, we try to find to true contour of an occluded object based on another one having the same shape. We have approximated the perspective projection to an affine transformation which is often used in literature according to the acquisition conditions. Let's consider the (c) im-

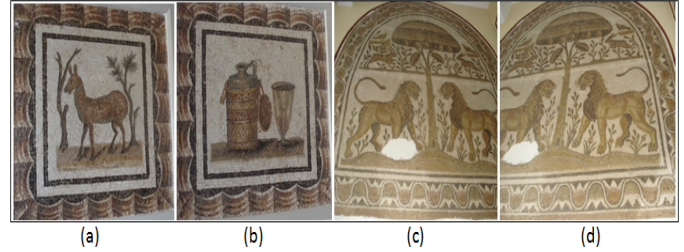


Fig. 7. Some mosaic images from the Bardo Museum of Tunisia.

age. As it's shown, this image contains many forms. Among all forms available in these two images, (c) and (d), taken from two sides, we use the Euclidean distance between the Affine Invariant Fourier Descriptors to localize two lions. The left one is partially occluded. We will use the right lion of the (d) image as template in order to have a better segmentation. In the last figure, images (a) and (b), we present in red

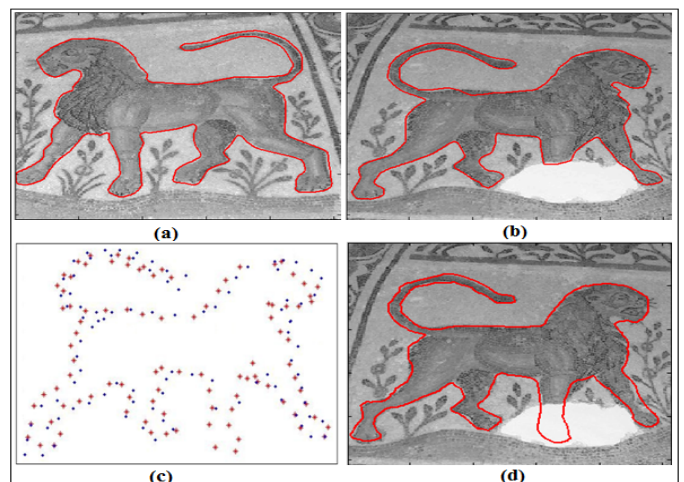


Fig. 8. Robust object detection in mosaic image.

the used curves to perform shape alignment and shape prior computation. The estimated values are $\alpha = 1$, $l_0 = -0.52$, $A = [-0.002, -0.003, 2.003, -0.35]$. We present by the (c) image the obtained curves alignment result and by the (d) image the segmentation result based on the proposed shape prior. Such a results is particularly interesting since it can be used for mosaic images restoration under partial occlusions and missing parts. The underlying idea is to extract similar forms using minimal distance between descriptors in order to define prior knowledge for such occluded or cluttered objects based on the presented affine shape alignment method for the purpose of a good segmentation result.

5. CONCLUSION

We presented in this paper an alternative approach to incorporate prior knowledge into a level set based active contours in order to have robust object detection in case of large shape distortion that can be analyzed by the class of affine transformations. We presented also a geometric solution to choose the reference shape in case of many available templates given that the statistical approach needs a training set and PCA. Then the application of a given classifier like Bayesian classifier to determine the appropriate reference shape like the work of Fang et Chan [25]. Given that the proposed approach invoke only pixels of the regions of variability between the shape of reference and the object of interest in the process of curve evolution and based on the fast Fourier transform for affine motion estimation and invariants computation, the method is faster compared to [15] and [26] where at each iteration shape descriptors are calculated for a given order that has to be set empirically. The obtained results are promising in the case of real and simulated data and the method can be used for the restoration of mosaic images in the archeological field. As future perspectives, we are working on integrating our model in the context of 3D object reconstruction from silhouettes sequence in order to refine the obtained 3D model.

6. REFERENCES

- [1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes : active contour models," *Int. J. of Comp.Vis.*, vol. 1, pp. 321–331, 1988.
- [2] S. Osher and J.A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulation," *J.of Computational Physics*, vol. 79, pp. 12–49, 1988.
- [3] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *Int. J. of Comp. Vis.*, vol. 22, pp. 61–79, 1997.
- [4] R. Malladi, J. Sethian, and B. Vemuri, "Shape modeling with front propagation : A level set approach," *PAMI*, vol. 17, pp. 158–175, 1995.
- [5] T. Chan and L. Vese, "Active contours without edges," *IEEE Trans. Imag. Proc.*, vol. 10, pp. 266–277, 2001.
- [6] M.A.Charmi, M.A.Mezghich, S.M'Hiri, S.Derrode, and F.Ghorbel, "Geometric shape prior to region-based active contours using fourier-based shape alignment," in *IST*, 2010, pp. 478–481.
- [7] S. Derrode M.A. Charmi and F. Ghorbel, "Using fourier-based shape alignment to add geometric prior to snakes," in *ICASSP*, 2009, pp. 1209–1212.
- [8] A. Foulonneau, P. Charbonnier, and F. Heitz, "Contraintes geometriques de formes pour les contours actifs orientes region : une approche basee sur les moments de legendre," *Traitement du signal*, vol. 21, pp. 109–127, 2004.
- [9] S. M'Hiri M.A. Mezghich, M. Sellami and F. Ghorbel, "Shape prior for an edge-based active contours using phase correlation," in *EUSIPCO*, 2013, pp. 1–5.
- [10] M. Leventon, E. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 316–323.
- [11] W. Fang and K.L. Chan, "Incorporating shape prior into geodesic active contours for detecting partially occluded object," *Pattern Recognition*, vol. 40, pp. 2163–2172, 2007.
- [12] Y. Chen, S. Thiruvenkadam, H.D. Tagare, F. Huang, D. Wilson, and E.A. Geiser, "On the incorporation of shape priors into geometric active contours," in *Proc. of IEEE Workshop on Variational and Level Set Methods in Computer Vision*, 2001, pp. 145–152.
- [13] P. Vandergheynst X. Bresson and J.P. Thiran, "A priori information in image segmentation : energy functional based on shape statistical model and image information," in *ICIP*, 2003, pp. 425–428.
- [14] V. Duay A. Allal N. Houhou, A. Lemkaddem and J.P. Thiran, "Shape prior based on statistical map for active contour segmentation," in *ICIP*, 2008, pp. 2284–2287.
- [15] A. Foulonneau, P. Charbonnier, and F. Heitz, "Affine-invariant geometric shape priors for region-based active contours," *PAMI*, vol. 8, pp. 1352–1357, 2008.
- [16] S. M'Hiri M.A. Mezghich and F. Ghorbel, "Fourier-based multi-references shape prior for region-based active contours," in *MELECON*, 2012, pp. 661–664.
- [17] S. M'Hiri M.A. Mezghich and F. Ghorbel, "Invariant shape prior knowledge for an edge-based active contours," in *VISAPP*, 2014, p. In Press.
- [18] F. Ghorbel, "Towards a unitary formulation for invariant image description: application to image coding," *An. of telecom.*, vol. 153, pp. 145–155, 1998.
- [19] F.Chaker and F.Ghorbel, "Apparent motion estimation using planar contours and fouriers descriptors," in *VISAPP*, 2010.
- [20] F.Ghorbel F.Chaker and M.T.Bannour, "Contour retrieval and matching by affine invariant fourier descriptors," in *IAPR*, 2007.
- [21] M.Saidani and F.Ghorbel, "Shape matching by affine movement estimation for 3d reconstruction," in *QCAV*, 2011.

- [22] M.Saidani and F.Ghorbel, "Affine invariant descriptors for 3d reconstruction of small archaeological objects," in *Compimage*, 2012.
- [23] F. Ghorbel, "Towards a unitary formulation for invariant image description : application to image coding," *An. of telecom.*, vol. 153, pp. 145–155, 1998.
- [24] L. Cohen, "On active contour models and balloons," *Graphical Models Image Process*, vol. 53, pp. 211–218, 1991.
- [25] W. Fang and K.L. Chan, "Using statistical shape priors in geodesic active contours for robust object detection," in *ICPR*, 2006, pp. 304–307.
- [26] S. Derrode M.A. Charmi and F. Ghorbel, "Fourier-based shape prior for snakes," *Pat. Recog. Let.*, vol. 29, pp. 897–904, 2008.

Segmentation of ancient documents in support of electronic file management

Ederson Marcos Sgarbi¹, Daniela de Freitas Guilhermino Trindade¹, Wellington Aparecido Della Mura¹ e Jacques Facon²

¹UENP – Universidade Estadual do Norte do Paraná – Centro de Ciências Tecnológicas
Bandeirantes – PR – Brasil

²PUCPR – Pontifícia Universidade Católica do Paraná – PPGIA
Curitiba – PR – Brasil

{sgarbi, danielaf, wellington}@uenp.edu.br, facon@ppgia.pucpr.br

Abstract — *This paper presents the segmentation of the background of ancient degraded documents over time based on mathematical morphology to color. The strategy of the proposed methodology is to apply morphological reconstruction operation for obtaining background of old images and thus the segmentation of the foreground color content. This method can be used in electronic document management in preprocessing for disposal of funds of old documents damaged tools. The tests showed good results applied to documents.*

Keywords - *Segmentation; Mathematical Morphology; Electronic Document Management;*

I. INTRODUCTION

In recent decades, the amount of electronic documents generated by companies soared, making the form of storage and demand [1]. With this increase there is a need to organize them digitally, facilitating the search faster. This control aims to increase productivity and promote knowledge management in an organization.

The ability of human beings to interact with organizational foundations became something prominent in our society. All the innovations make difficult to control the management of information and knowledge. Focusing on this kind of management used by many agencies and companies, we have adopted the EDM (Electronic Document Manager) strategy that gives for the users security and sustainability in the information store.

The use of solutions for optical character recognition have been widely used for various purposes to manage electronic documents. However, when working with complex, damaged or low quality scanned documents, it is needed to develop more complex solutions.

The pages of ancient documents suffer serious deteriorations caused by inappropriate storage and time. To solve these problems, a strategy consists in segmenting the background and foreground of each page, storing only relevant information, facilitating the indexing of documents for future electronic management.

We can find in the literature some research focused on to old document processing. The Granado, Pina and Muge's approach [2] performs the old document segmentation by binary mathematical morphology. The Baird's approach [3] uses training region-based algorithms with ground truth images to segment color documents. The Journet, Ramel, Mulot's approach [4] presents a color texture page analysis of old documents through the detection of specific areas containing guidelines and frequencies, with colorful pictures.

In [5] the paper presents a recovery process that performs the background and foreground segmentation based on color mathematical morphology. This recovery process can be the initial step of electronic document management.

Figure 1 illustrates the general flowchart of scanned documents management approach.

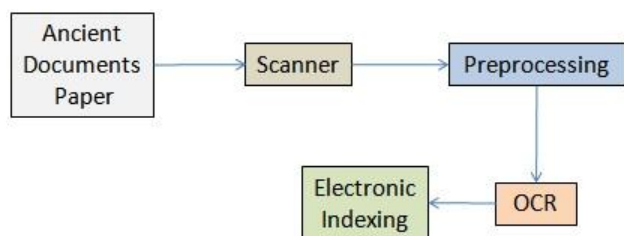


Figure 1: Flowchart of electronic file management.

II. PROPOSED METHODOLOGY

The strategy used to help the Management of Electronic Documents in character recognition and indexing process is based on background and foreground segmentation form color mathematical morphology [5].

A. Electronic Document Management

Vital and strategic information in any organization, regardless of the economic activity, needs to be preserved, disseminated and quickly accessed.

The quantity of information and documents grows up every day increasingly larger, and developing new tools is essential. The EDM (Electronic Document Manager) is the more useful tool for being humans processing and recording information. EDM also aims simultaneously to reduce the physical space and access documents.

Information and Communication Technologies (ICTs) allow generating, store and retrieving documents, being called, for this reason, electronic documents. Thus the objective of the proposed strategy is making easier the elimination of damaged backgrounds and facilitating access to important document contents.

Solutions for optical character recognition have been widely used for various purposes. Commercial OCRs successfully perform character recognition when the documents are simple without degradations. However, when working with old and / or damaged documents, the character recognition is not always well succeeded.

B. Ancient Documents

A document may contain text, pictures, drawings and graphics, which may prove the existence of a fact, the accuracy or truth of a statement. Historical documents are the historic sources of contexts, facts, arguments which justify any historical event. These documents are considered old. Very often, these ancient documents suffer greatly from the deterioration of the paper background caused by improper storage and time. A strategy to tackle this problem is the preprocessing.

C. Preprocessing

This step consists in processing the deteriorated page images using color mathematical morphology with morphological reconstruction operation. This preprocessing is based on uncorrelated color space and lexicographic order [5]. Figure 2 illustrates the flowchart for this step.

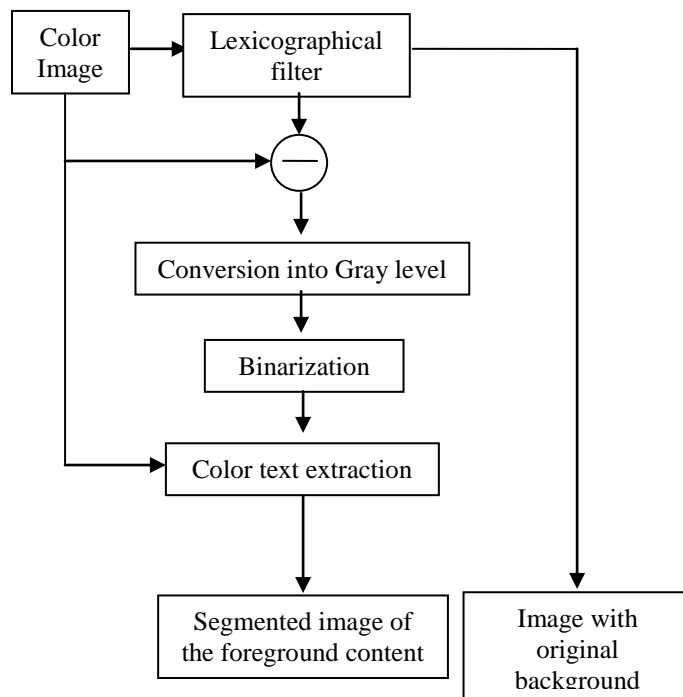


Figure 2: Flowchart of the preprocessing of old documents [5].

- *Mathematical Morphology*

Mathematical morphology is a non linear theory to process geometrical structures in digital images [2]. The basic principle of mathematical morphology is extracting information on the geometry and content of images using two basic operators, erosion and dilation [6]. The mathematical morphology, originally developed for binary images has been largely extended to grayscale images. In grayscale morphology, the two operators, erosion and dilation are based on the concepts of minimum and maximum gray levels that do not present any computational difficulty.

New horizons focus on the extension to color images [7]. The challenge in color mathematical morphology is to define concepts of minimum and maximum with colors. How can we sort colors? Mathematical concepts of minimum and maximum and theoretical challenges reside in the appropriate choice of color space and the definition of the ordering of colors. Multivariate data can be sorted by various techniques, for instance complete, marginal, conditional or lexicographic orders [8]. The model adopted for the ordering colors is important, since color images may be represented by several color models [9].

- *Color Space*

A color space is a mathematical model used to represent a range of colors. Human color vision is a complex process that is still not completely understood.

Color perception by humans is a psycho-physiological phenomenon that is still not completely understood. However, the physical nature of the colors can be expressed as a formal basis supported by experimental and theoretical results. There are several color spaces such as RGB, HSI, HSV, YCrCb, Luv, Lav, XYZ and others. In image processing area, RGB and HSI color spaces are more widely used. Due to correlation between components, RGB color space has the disadvantage of generating false colors. Since the HSI and YCrCb color spaces are more appropriate to image processing area because their channels are uncorrelated. So in this work we use the uncorrelated HSI color space

[10]. But we can notice that there are few studies using YIQ and YCrCb color spaces [5] [11].

- *Lexicographic Order*

It is a sorting technique which refers to the order in which the words are arranged in dictionaries. In [7] 10] and [12], it is noticed that the lexicographic order is the most used in color mathematical morphology in color. In this first sort order is decided by a channel, followed by a second channel and a third channel after. If the first comparison does not satisfy the components need to perform the second comparison and so on Equation 01 illustrates the generic lexicographic order to rank the vectors with the three color channels. With the lexicographical ordering is necessary to define the channel order, with space for a given color combinations thereof. Equation 01 shows the generic lexicographic order to rank the vectors with three color channels. Let be a pixel $P(x_1, y_1, z_1)$ and a neighbor pixel $P_2(x_2, y_2, z_2)$. One can define the infimum between the colors of $P(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ as follows:

$$P(x_1, y_1, z_1) < P_2(x_2, y_2, z_2) \Leftrightarrow \begin{cases} x_1 < x_2 \\ or \\ y_1 < y_2 \text{ if } x_1 = x_2 \\ or \\ z_1 < z_2 \text{ if } y_1 = y_2 \text{ and } x_1 = x_2 \end{cases} \tag{01}$$

- *Channel Order*

After choosing a color space and the color order, it is necessary to choose the color space channel order before using the morphological operators. For example, with lexicographic order and HSI color space, there are the possible color space channel orders: HSI, ISH, IHS, SHI, SIH and HSI.

- *Color Reconstruction*

Color reconstruction is a powerful operator. The reconstruction function $\rho(f)$ from geodesic dilation $\delta_g(f)$ is defined as [5]:

$$\rho(f) = \lim_{n \rightarrow +\infty} \underbrace{\delta_g(\delta_g(\dots\delta_g(f)))}_n \text{ with } \delta_g(f) = \delta(f) \wedge g \quad (03)$$

where the color geodesic dilation $\delta_g(f)$ is based on restricting the color dilation of the marker image f (image to be reconstructed) to the mask g (model).

To perform the reconstruction process, is necessary to follow the following steps:

- ✓ First step: Define the marker image and mask (model).
- ✓ Second step: Apply the geodesic dilation to the marker image;
- ✓ Third stage: Compare the dilated marker image with the mask image, until convergence

- *Binarization*

The binarization consists in separating foreground from background, ie separating the text from the deteriorated image background . Several algorithms have been studied and tested, and we have empirically concluded that the best binarization algorithm was the Sauvola's technique [13].

D. Tesseract-OCR

It is an engine of open source OCR for optical character recognition devised by Smith [14], developed by Hewlett-Packard using the language C++.

The character recognition process is performed according to the document will be processed. When the documents are ancient, deteriorated background and yellowish paper degradations decrease the performance of commercial OCR. So pre-processing for

eliminating these problems described as before are necessary [15] to increase the performance of Tesseract-OCR .

E. Indexing

Using the Tesseract-OCR makes possible the indexation of ancient documents based on character recognition. In this case, it is necessary to define the layouts of old documents to perform a efficient recognition and indexation.

F. Evaluation Metrics

The metrics are used to measure the efficiency of segmentation methods [5]. The most commonly used segmentation metrics are described in [15, 16, 17].

III. EXPERIMENTS

The experiments were carried out using a database of 100 ancient document images appearing with digitization errors, foxed background and bleed-through effects.

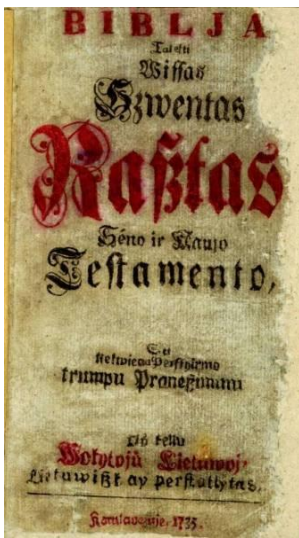
Experimental results carried onto these 100 ancient document images have proven that our approach is visually effective to recover ancient texts in uneven and foxed background images.

To perform a numerical evaluation of our approach, some ground-truth images with the content of the foreground were prepared.

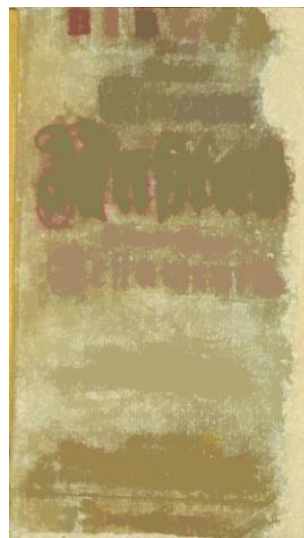
The F Measure and Negative Rate Metric NRM metrics used to measure the efficiency of our approach have obtained 99.29% and 0,12% of accuracy and error (respectively) in segmenting the foreground content.

The Figure 3 illustrates the background removal pre-processing and result for an old image.

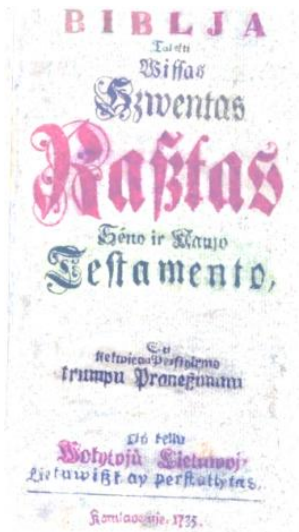
Note on the results that the fund was well reconstructed using the reconstruction operation in color. The segmentation of foreground content was also obtained.



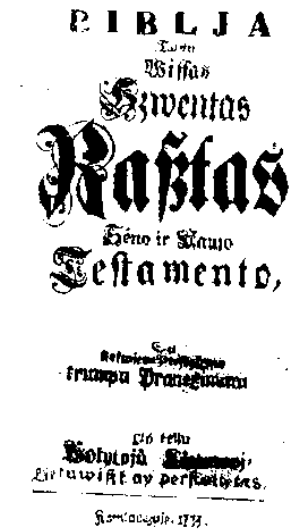
a) Original Image



b) Reconstruction of background by color reconstruction with ISH order.



c) Subtracted image (a) and (b)



d) Binarized image by Sauvola



e) Segmented image with foreground content

Figure 3: Example of old document pre-processing.

Figure 4 illustrates the background removal pre-processing and result for an more complex image containing pictures, yellow background, printed text in two columns.



a) Original Image



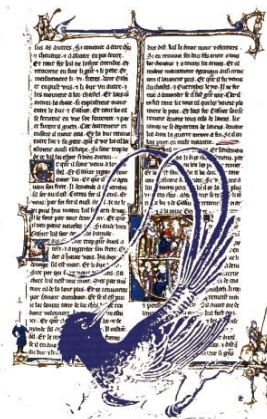
b) Reconstruction of background by color reconstruction with ISH order.



c) Subtracted image (a) and (b)



d) Binarized image by Sauvola



f) Segmented image with foreground content

Figure 4: Example of old document pre-processing.

We may notice that the results depicted in Figures 2 and 3 show some foreground pixels not correctly targeted. The problem lies in the choice of the thresholding technique. It happens because some foreground pixels were not binarized as expected. After testing several binarization techniques as Johannsen, MCC, Otsu and Sauvola's algorithm, we have concluded that Sauvola technique has showed be the best.

IV. CONCLUSIONS

Electronic managers require technical documents to be implemented to make easier and add new technologies to enhance existing electronic managers, providing more speed and agility in the storage and indexation of documents.

In this paper, it was shown that the innovative approach to perform the background and foreground segmentation of ancient document images using morphological color operators presented by [5] was useful and efficient for this proposal.

This application can be also used for several different business areas such as: libraries, registries, companies, governments, all kinds of entities that need to store and make available old documents.

Future work resides in the idea to implement a friendly graphical interface for users for electronically managing old document images deteriorated by digitization errors, uneven background and bleed-through effects.

REFERENCE

- [1] DELLA MURA, W; PEPIS, L; YASUI, P; POZZA, R; BASTIANI, J A. Implementando um gestor eletrônico de documentos e informações. In: 4ª Conferência Ibérica de Sistemas e Tecnologias de Informação - Cisti'2009, Póvoa de Varzim – Portugal.
- [2] GRANADO, I.; PINA, P.; MUGE, F. "Automatic feature extraction on pages of antique books through a mathematical morphology based methodology", 10 Encontro Português de Computação Gráfica, 2001.
- [3] BAIRD, H. S. et al. Document image content inventories. Proc. IST/SPIE Document Recognition and Retrieval XIV Conf., San Jose, CA, 28 January - 1 February, 2007.
- [4] JOURNET, N.; RAMEL, J. Y.; MULLOT, R. Document image characterization using a multiresolution analysis of the texture: application to old documents. Spring-Verlag - IJDAR 2008 International Journal on Analysis and Recognition, v. 1, p. 918, 2008.
- [5] SGARBI, E.M; DELLA MURA, W.A.; MOYA, N; FACON, J.; AYALA, H. L.; "Restoration of Old Document Images using different color spaces", VISAPP 2014 9th International Joint Conference on Computer Vision, 2014, Lisbon, Portugal. Vol. 1. p. 82-88.
- [6] FACON, J. "Mathematical Morphology: theory and practice" (In Portuguese) , Editor Facon Jacques, 1996.
- [7] APTOULA, E.; LEFÉVRE, S. A comparative study on multivariate mathematical morphology. Preprint submitted to Elsevier Science, v. 1, n. 1, p. 1-37, 2007.
- [8] TRAHANIAS, P.; VENETSANOPOULOS, A. Colour edge detectors based on multivariate ordering. In: Proceedings of SPIE, Visual Communications and Image Processing'92, Petros Maragos, University of Toronto, v. 1818, p. 1396-1407, 1992.
- [9] SANGWINE, S.; HORNE, R. The colour image processing handbook. London Chapman Hall, 1998.
- [10] ORTIZ, F. et al. Colour mathematical morphology for neural image analysis. Real-Time Imaging 8 Elsevier Science Ltd, v. 8, p. 455{465, 2002.
- [11] POPOV, A. T. Fuzzy mathematical morphology and its applications to colour image processing. 15th Internacional Conference on Computer Graphics, Visualization e Computer Vision 2007, 2007.
- [12] PITAS, I.; VENETSANOPOULOS, A. N. Order statistics in digital image processing.

Proceedings of the IEEE, v. 80, n. 12, p. 1893{1923, 1992.

- [13] SHAFAIT, F.; KEYSERS, D.; BREUEL, T. M. Efficient implementation of local adaptive thresholding techniques Proceedings of the 15th Document Recognition and Retrieval Conference (DRR-2008), v. 6815, 2008.

- [14] SMITH, R. Na overview of the Tesseract OCR engine. Int. Conf. on document Analysis and Recognition (ICDAR), Curitiba, Brazil, 2007.
- [15] SEZGIN, M.; SANKUR, B. Selection of thresholding methods for non destructive testing application. ICIP2001, v. 3, p. 764 - 767, 2001.
- [16] BIMBO, A. D. Visual information retrieval. Morgan Kaufmann, 1999.

- [17] GATOS, B.; NITIROGIANNIS, K.; PRATIKAKIS, I. Document image binarization contest. ICDAR - International Conference on Document Analysis and Recognition, p.1375 - 1382, 2009.

Distances and Kernels Based on Cumulative Distribution Functions

Hongjun Su and Hong Zhang*

Department of Computer Science and Information Technology

Armstrong Atlantic State University

Savannah, GA 31419 USA

E-mail: hongjun.su@cs.armstrong.edu, hong.zhang@armstrong.edu

**contact author*

Conference: IPCV'14

Keywords: Cumulative Distribution Function, Distance, Kernel, Similarity

Abstract

Similarity and dissimilarity measures such as kernels and distances are key components of classification and clustering algorithms. We propose a novel technique to construct distances and kernel functions between probability distributions based on cumulative distribution functions. The proposed distance measures incorporate global discriminating information and can be computed efficiently.

1. Introduction

The kernel is a similarity measure that is the key component of support vector machine ([4]) and other machine learning techniques. More generally, a distance (a metric) is a function that represents the dissimilarity between objects.

In many pattern classification and clustering applications, it is useful to measure the similarity between probability distributions. A large number of divergence and affinity measures on distributions has already been defined in traditional statistics. These measures are typically based on the probability density functions and are not effective in detecting global changes.

In this paper, we propose a family of distances and kernels that are defined on the cumulative distribution functions, instead of densities.

This paper is organized as follows. Section 2 introduces kernels and distances commonly defined on probability distributions. In Section 3, a new family of distance and kernel functions based on cumulative distribution functions is proposed. Experimental results on Gaussian

mixture distributions are presented in Section 4. In Section 5 we provide conclusions and future works.

2. Distance and Similarity Measures Between Distributions

Given two probability distributions, there are well known measures for the differences or similarities between the two distributions.

The Bhattacharyya affinity ([1]) is a measure of similarity between two distributions:

$$B(p, q) = \int \sqrt{p(x)q(x)} dx$$

In [7], the probability product kernel is defined as a generalization of Bhattacharyya affinity:

$$k^{prob}(p, q) = \int p(x)^{\rho} q(x)^{\rho} dx$$

The Bhattacharyya distance is a dissimilarity measure related to the Bhattacharyya affinity:

$$D_B(p, q) = -\ln\left(\int \sqrt{p(x)q(x)} dx\right)$$

The Hellinger distance ([6]) is another metric on distributions:

$$D_H(p, q) = \sqrt{\frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx}$$

The Kullback-Leibler divergence ([8]) is defined as:

$$D_{KL}(p, q) = \int \left(\ln \frac{p(x)}{q(x)} \right) p(x) dx$$

All these similarity/dissimilarity measures are based on the point-wise comparisons of the probability density functions. As a result, they are inherently local comparison measures of the density functions. They perform well on smooth, Gaussian-like distributions. However, on discrete and multimodal distributions, they may not reflect the similarities and can be sensitive to noises and small perturbations in data.

Example. Let p be the simple discrete distribution with a single point mass at the origin and q the perturbed version with the mass shifted by a . (Figure 1)

$$p(x) = \delta(x)$$

$$q(x) = \delta(x - a)$$

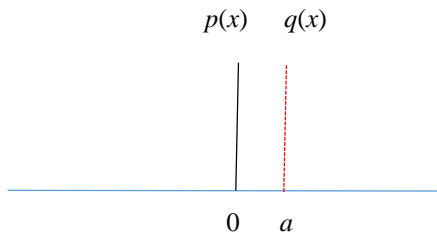


Figure 1. Distributions

The Bhattacharyya affinity and divergence values are easy to calculate:

$$B(p, q) = \int \sqrt{p(x)q(x)} dx = 0$$

$$D_B(p, q) = -\ln \left(\int \sqrt{p(x)q(x)} dx \right) = \infty$$

$$D_H(p, q) = \sqrt{\frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx} = \infty$$

$$D_{KL}(p, q) = \int \left(\ln \frac{p(x)}{q(x)} \right) p(x) dx = \infty$$

All these values are independent of a . They indicate minimal similarity and maximal dissimilarity.

The earth mover's distance (EMD), also known as the Wasserstein metric ([3]), is defined as

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int d(x, y)^p d\gamma \right)^{1/p}$$

where $\Gamma(\mu, \nu)$ denotes the set of all couplings of μ and ν . EMD does measure the global movements between the distributions. However, the computation of EMD involves solving optimization problems and is much more complex than the density based divergence measures.

Related to the distance measures are the statistical tests to determine if two samples are drawn from different distributions. Examples of such tests include the Kolmogorov-Smirnov statistic ([10]) and the kernel based tests ([5]).

3. Distances on Cumulative Distribution Functions

A cumulative distribution function (CDF) of a random variable X is defined as

$$F(x) = P(X < x)$$

Let F and G be CDFs for the random variables with bounded ranges (i.e. their density functions have bounded supports). For $p \geq 1$, we define the distance between the CDFs as

$$d_p(F, G) = \left(\int |F(x) - G(x)|^p dx \right)^{1/p}$$

It is easy to verify that $d_p(F, G)$ is a metric. It is symmetric and satisfies the triangle inequality. Because CDFs are left-continuous, $d_p(F, G) = 0$ implies that $F = G$.

When $p = 2$, a kernel can be derived from the distance $d_2(F, G)$:

$$k(F, G) = e^{-\alpha d_2(F, G)^2}$$

To show that k is indeed a kernel, consider a kernel matrix $M = [k(F_i, F_j)]$, $1 \leq i, j \leq n$. Let $[a, b]$ be

a finite interval that covers the support of all density functions $p_i(x), 1 \leq i \leq n$.

$$d_2(F_i, F_j) = \left(\int_a^b |F_i(x) - F_j(x)|^2 dx \right)^{1/2}$$

This metric is induced from the norm of the Hilbert space $L^2([a, b])$. Consequently the kernel matrix M is positive semi-definite, since it is the kernel matrix of the Gaussian kernel for $L^2([a, b])$. Therefore, k is a kernel.

Remarks. The formula for $d_p(F, G)$ resembles the metric induced by the norm in $L^p(\mathbf{R})$. However a CDF F cannot be an element of $L^p(\mathbf{R})$ because $\lim_{x \rightarrow \infty} F(x) = 1$. The condition of bounded support will guarantee the convergence of the integral. In practical applications, this will not likely be a limitation. Theoretically the integral could be divergent without this constraint. For example, let F be the step function at 0 and $G(x) = x/(x+1), x \geq 0$. Then

$$d_1(F, G) = \int_0^\infty \frac{1}{x+1} dx = \infty$$

Given a data sample, (X_1, X_2, \dots, X_n) , an empirical CDF can be constructed as:

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n I_{X_k < x}$$

which can be used to approximate the distance $d_p(F, G)$.

When $p = \infty$, we have

$$d_\infty(F, G) = \max_x |F(x) - G(x)|$$

The distance d_∞ is similar to the Kolmogorov-Smirnov statistic ([10]).

Example. Consider the same example as in the previous section. The CDFs are illustrated in Figure 2.

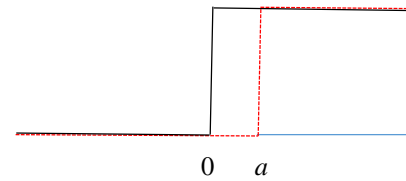


Figure 2. CDFs

The proposed distance function has the value:

$$d_p(F, G) = \left(\int_0^a 1 dx \right)^{1/p} = a^{1/p}$$

For $p < \infty$, the distance value is dependent on a . The kernel value is:

$$k(F, G) = e^{-ad_2(F, G)^2} = e^{-a\sqrt{a}}$$

The computation of the distance $d_p(F, G)$ is straightforward. For a discrete dataset of size n , the complexity for computing the distance is $O(n)$. On the other hand, the computation of earth mover's distance requires the Hungarian algorithm ([9]) with a complexity of $O(n^3)$.

4. Experimental Results and Discussions

The CDF based kernels and distances can be effective on continuous distributions as well.

A Gaussian mixture distribution ([2]) and its variations, shown in Figure 3, are used to test the kernel functions. The first chart shows the original Gaussian mixture. The other two distributions are obtained by moving the middle mode. Clearly the second distribution is much closer to the original distribution than the third one.

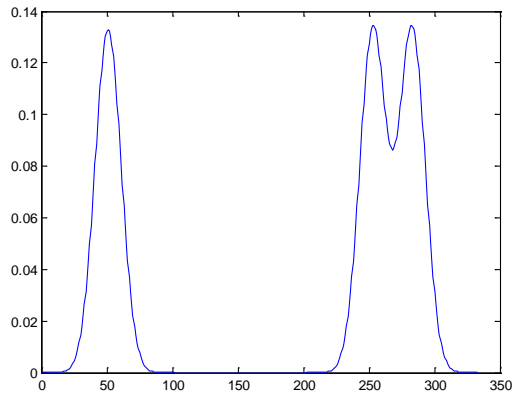
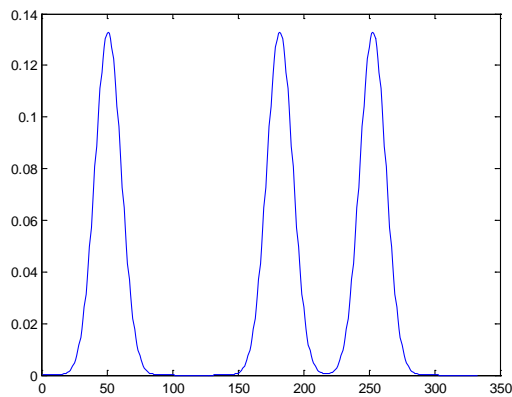
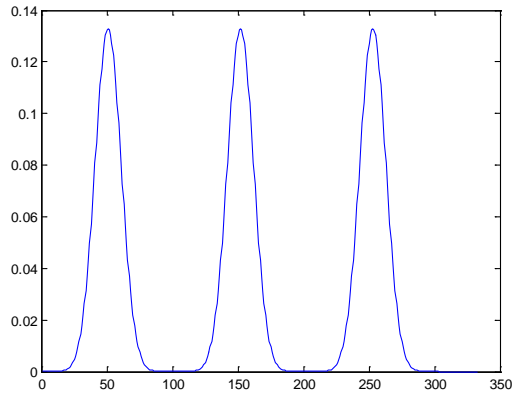


Figure 3. A Gaussian mixture and variations

Indexed in the same order as Figure 3, the Bhattacharyya kernel matrix for the three distributions is:

$$\begin{pmatrix} 1 & 0.775 & 0.715 \\ 0.775 & 1 & 0.715 \\ 0.715 & 0.715 & 1 \end{pmatrix}$$

The Bhattacharyya kernel did not clearly distinguish the second and the third distributions when comparing to the original. There is no significant difference between the kernel values k_{12} and k_{13} , which measure the similarities between the original distribution and the two varied distributions.

The kernel matrix of our proposed kernel is:

$$\begin{pmatrix} 1 & 0.123 & 1.67 \times 10^{-6} \\ 0.123 & 1 & 4.68 \times 10^{-5} \\ 1.67 \times 10^{-6} & 4.68 \times 10^{-5} & 1 \end{pmatrix}$$

The CDF based kernel showed much greater performance in this example. The kernel values k_{12} and k_{13} clearly reflect the larger deviation (less similarity) of the third distribution from the original.

This is due to the fact that the density based Bhattacharyya kernel does not capture the global variations. The CDF based kernel is much more effective in detecting global changes in the distributions.

5. Conclusions and Future Work

In this paper, we presented a new family of distance and kernel functions on probability distributions based on the cumulative distribution functions. The distance function was shown to be a metric and the kernel function was shown to be a positive definite kernel. Compared to the traditional density based divergence functions, our proposed distance measures are more effective in detecting global discrepancy in distributions. Experimental results on generated distributions were discussed.

This method can be extended to high dimensional distributions. The advantages of CDF can be maintained in the high dimensional cases. However, there will be significant cost in directly computing high dimensional CDFs. We plan to investigate specifically the 2D extension which could yield useful results for image processing.

Acknowledgment. The authors wish to thank the referees for their extremely helpful comments and suggestions.

6. References

- [1] Bhattacharyya, A., "On a measure of divergence between two statistical populations defined by their probability distributions". Bulletin of the Calcutta Mathematical Society 35: 99–109, (1943).
- [2] Bishop, Christopher, Pattern recognition and *machine learning*. New York: Springer, (2006).
- [3] Bogachev, V.I.; Kolesnikov, A.V. "The Monge-Kantorovich problem: achievements, connections, and perspectives". Russian Math. Surveys 67: 785–890.
- [4] Boser, B. E.; Guyon, I. M.; Vapnik, V. N., "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. p. 144, (1992).
- [5] Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Scholkopf, B.; Smola, A., "A kernel two-sample test", J. Machine Learning Research, 13, 723-773, (2012).
- [6] Hellinger, Ernst, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen", Journal für die reine und angewandte Mathematik (in German) 136: 210–271, (1909).
- [7] Jebara, T.; Kondor, R.; Howard, A., "Probability Product Kernels," J. Machine Learning Research, 5, 819-844, (2004).
- [8] Kullback, S.; Leibler, R.A., "On Information and Sufficiency". Annals of Mathematical Statistics 22 (1): 79–86, (1951).
- [9] Munkres, J., "Algorithms for the Assignment and Transportation Problems", Journal of the Society for Industrial and Applied Mathematics, 5(1):32–38, (1957).
- [10] Smirnov, N.V., "Approximate distribution laws for random variables, constructed from empirical data" Uspekhi Mat. Nauk , 10 pp. 179–206 (In Russian), (1944).

A Spot Matching Method for Landmark Matched Pairs using the Second Neighbor Spots in 2-DE

Chan-Myeong Han¹, Dae-Seong Jeoune², Yun-Kyoo Ryoo³, Sung-Woo Han⁴, Hwi-Won Kim⁵,
Wookhyun Kim⁶, and Young-Woo Yoon⁶

¹M+VISION Co. Ltd., Daegu Metropolitan City, Republic of Korea

²Department of Media Design, Daegu Future College, Kyungsan, Kyungbuk, Republic of Korea

³Department of Medical Computer Science, Daegu Health College, Daegu, Republic of Korea

⁴Department of Computer Engineering, Daegu Science University, Daegu, Republic of Korea

⁵Education Center for IT, Kyeongbuk College, Yeungju, Kyungbuk, Republic of Korea

⁶Department of Computer Engineering, Yeungnam University, Kyungsan, Kyungbuk, Republic of Korea

Abstract - In 2D gel spot matching, landmark spots mean highly distinguishable spots in color, size and shape. Some methods need matched pairs of landmark spots from the very start and they play a very key role in obtaining the right matching results. Researchers manually detect matched pairs populated evenly all around the gel image for this reason. However, manual detection of the matched pairs is the main obstacle in minimizing researcher's intervention and implementing a fully automated spot matching process. In this paper, we propose a fully automated and highly reliable process of detecting landmark matched pairs using topological patterns of the second neighbor spots. This method can produce a number of precise landmark matched pairs enough with less efforts and it can benefit many spot matching methods based on matched pairs of landmark spots in that they can use more reliable and sufficient landmark matched spots with good quality.

Keywords: Spot matching, landmark matched spot, neighbor spot, 5-NNG, grassfire spot matching

1 Introduction

It is found that whole genome sequence cannot explain life phenomena enough and has a lot of limitation to find disease related genes after successful genome sequencing of over 40 species[1]. Studies on proteins and interactions among them are considered as one of key fields because genes are expressed into proteins through mRNAs. Proteomics is the large-scale study of learning functions of proteins and the very basic process is to identify proteins included in cells. The two-dimensional electrophoresis(2-DE) is the most frequently used method in proteomics[2, 3].

In the study of proteins, spot matching is a main bottleneck and the implementation of fast and precise spot matching algorithm without researcher's intervention is the most essential part to upgrade proteomics one level up[4]. However, it is impossible to obtain the same gel image each time due to many experimental parameters even if the same

sample is used for a couple of electrophoresis experiments. It makes spot matching more difficult.

Many spot matching algorithms have been proposed for several decades. Among them, there are some methods for spot matching which require matched pairs of landmark spots as input. They try to match spots around the pre-given landmark matched pairs[6, 7]. Other methods evaluate parameters for transformation using landmark matched pairs and try to match spots after transforming one of two 2-DE gel images[8, 9]. Only one matched pair of spots is needed as a seed pair and it assumes that the pair should be correct one in any case so that it may work properly[9].

In this paper, a new method for detecting matched pairs of landmark spots is proposed. Matched pairs of landmark spots should be very accurate because the wrong pairs lead to wrong results without fail. Matched pairs are detected using topological patterns of neighbor spots based on the literature [10]. However, the more spots are densely populated, the more similar topological patterns could occur and it increases false positive matching results. The proposed method introduces topological patterns of the second neighbor spots, while the neighbor spots of the first neighbors to reduce false positive results. It can be used directly in acquiring seed spot pairs in grassfire spot matching algorithm[9]. It can be also applied as an automated method for matched pairs of landmark spots for any other algorithms based on landmark spots.

2 Materials and methods

2.1 Definition of spot matching problem

Spot matching starts with two sets of 2D points, $P=\{p_1, p_2, \dots, p_m\}$ and $Q=\{q_1, q_2, \dots, q_n\}$ where centroids of spots from reference gel image $p_i=(x_i, y_i)$ and centroids of spots from target gel image $q_j=(x_j, y_j)$. Here, spot matching is to find the maximum set of one to one matching pairs between P and Q , $M=\{(p_{i1}, q_{j1}), (p_{i2}, q_{j2}), \dots, (p_{il}, q_{jl})\}$, where $p_{il} \in P$, $q_{jl} \in Q$, and $l \leq \min(m, n)$.

There might be missing spots in some cases. The word “missing” means that a spot really exists but it is not detected in the process of spot detection for many reasons. There might also be outlier spots in some cases. “Outlier” means that the counterpart of a spot does not exist by bio-chemical reasons such as diseases or environmental conditions. However, the two words are used as the same meaning. Both cases cause spots with no counterpart.

2.2 Definition of the first and the second neighbor spots using graphs

Neighbor spots are defined as spots located near around a certain spot, which is a very ambiguous definition. Therefore, a graph theory is used to define neighbor spots mathematically for a clear definition of neighbor spot[10]. Edges are formed between two spots when a graph theory is applied in a set of points. The two spots are in the relationship of neighbors each other when they have a common edge. One spot usually has many neighbor spots and the number of neighbor spots depends on a graph theory to be applied and the topology of the whole points. We define neighbor spots directly connected to a spot as the first neighbor spots. We introduce a new definition of the second neighbor spot as neighbor spots of the first neighbor spots for the central spot. The example for the first neighbor spots and the second neighbor spots is shown in the Figure 1. The equations (1) and (2) are the mathematical notations for the first and the second neighbor spots.

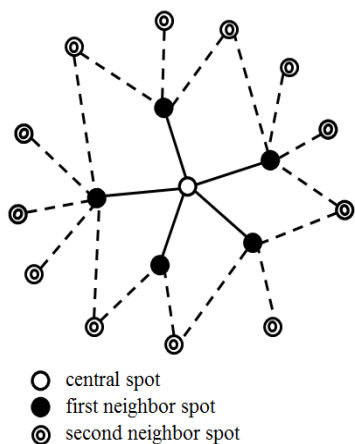


Figure 1. The first and the second neighbor spots for the central spot

$$N_{1,G}(v) = \{u \mid vu \in E\} \tag{1}$$

$$N_{2,G}(v) = \{N_{1,G}(u) \mid vu \in E\} \tag{2}$$

In the equation (1), The notation ‘N’ means first letter of the word “neighbor” and the subscripts ‘I’ and ‘G’ denote the first neighbor spot and a graph to be applied, respectively. The notations ‘v’ and ‘u’ are nodes of the graph and ‘E’ is a

set of edges in the graph. In the equation (2), the subscript ‘2’ means the second neighbor spot in the same way as the equation (1). In the Figure 1, there is clear classification between the first neighbor spots and the second ones. The first neighbor spots might be second neighbor spots at the same time because edges are connected to one another intricately in the most cases.

2.3 Spot matching method using topological patterns of the first neighbor spots

The method in the literature [10] tried to match spots by measuring the similarity of two topological patterns for first neighbor spots. It applies a certain graph to two sets of spots which are called reference and target gel in order to define neighbor spots. It limits the scope of comparison to neighbor spots $N_{1,G}(p_i)$ and $N_{1,G}(q_j)$ as in the Figure (2). The reason why it takes only the first neighbor spots is that 2-DE gel images have unpredictable distortion all the time both globally and locally. The larger the scope of comparison is, the more distortion affects the matching process, which leads to a failure in spot matching.

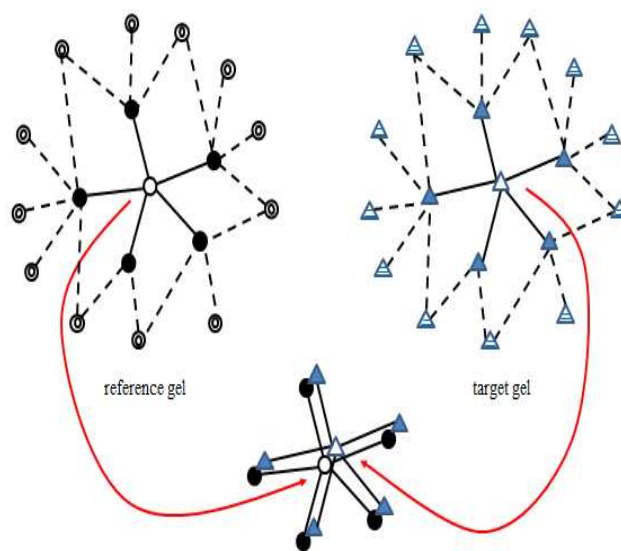


Figure 2. Spot matching by topological patterns of the first neighbor spots

Each set of the first neighbor spots forms a topological pattern and these two topological patterns are evaluated on how much they are similar. The similarity measurement is performed using similarity transform and the normalized Hausdorff distance. It yields three factors such as the number of matched pairs for the first neighbor spots, the number of unmatched spots(outlier spots) and the normalized Hausdorff distance. All of the spots from reference gel are tested to search a corresponding spot in target gel. The best result for every possible combination of pairs (p_i, q_j) is accepted to determine a matched pair for p_i in the test. A pair with the

biggest number of matched pairs, the smallest number of unmatched spots and the shortest normalized Hausdorff distance can be thought as the best result or the best match.

2.4 Detection of landmark matched pairs by the second neighbor spots

Some spot matching algorithms require some correctly matched pairs of spots as input[5-8]. These pairs are usually used for transforming gel image in the level of image or point pattern to match the other spots. For an unique example, the grassfire spot matching method requires only one matched pair as a seed spot where matching process starts[9]. Almost all of the methods use pairs which researchers manually search for accuracy because the pairs play a very crucial role in spot matching process. It is convenient to target landmark spots when they manually try to match spots. Landmark spots can be considered as a spot highly distinguishable in color, size and shape. It is a very intuitive and heuristic way suitable for humans.

In algorithm levels, the word “landmark” could differ from meaning by human because some spot matching algorithms can notice other facts other than color, size and shape which humans cannot intuitively detect. The broad meaning of “landmark” could be defined as spots with some traits easy to search corresponding spots. In this paper, we propose a highly reliable method for fully automated detection of matched pairs of landmark spots whose topologies are well preserved against distortion.

In 2-DE, a lot of variations in the digitized gel image are involved by experimental environments. The trait of these variations is known to be uncontrollable by all means. The same results cannot be achieved even from two consecutive cases of experiments conducted in the same laboratory with the exactly same experimental tools and conditions. It always does not bring the correct result to distinguish spots by color, size and shape using computer programs for this reason.

We noted that topological patterns of spots are more robust than color, size and shape of spot against the inherent trait of variation in 2-DE and it made us to use spot matching algorithm proposed in the literature [10]. It is a kind of spot matching algorithm and determines matched pairs of spots by similarity measurement for topological patterns of the first neighbor spots. It produces information in the process of matching spots such as the number of matched pairs of neighbor spots, the number of unmatched spots and the normalized Hausdorff distance. The spot pairs with high score of the information can be classified as matched pairs of landmark spots among randomly combined pairs in the process of spot matching. The matched pairs with well conserved topologies are regarded as so-called “landmark” in this paper.

In this manner, detecting matched pairs can be fully automated in a programmatic way rather than in manual. However, the problem is that the ratio of false-positive results is increased dramatically as the density of spots in gel images increases. It is because probability on the occurrence of

similar topological patterns increases in the case of densely populated gel image and it hamper the accuracy of spot matching. We make up for the drawback by introducing the second neighbor spots. As mentioned above, the second neighbor spots are the neighbor spots of the first neighbor spots. The false-positive results can be reduced by using the second neighbor spots in which a broader range of neighbor spots are included to form more complex topological patterns of neighbor spots.

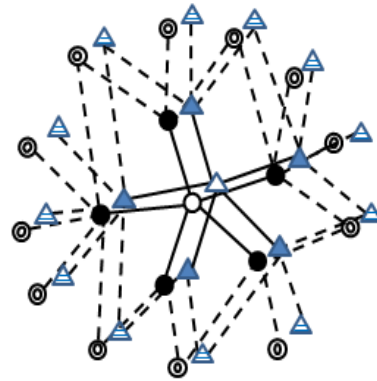


Figure 3. Spot matching by topological patterns of the second neighbor spots

A 5-NNG(nearest neighbor graph with degree 5) is applied to form edges among spots and the first neighbor spots are used in spot matching in the literature [10]. We modify it by using the second neighbor spots instead of the first ones as in the Figure 3. We only get top 10% of the matched pairs as candidates of landmark matched pairs. The matched pairs are scored according to the three factors as follows. “The bigger the number of matched pairs and the smaller the number of unmatched spots” means that it is more likely to be the right match. “The shorter the normalized Hausdorff distance” presents more similar patterns.

3 Experiment and result

3.1 Data set for experiment

Manually confirmed matched pairs of landmark spots is required to verify the propose method. However it takes a lot of effort and cost because a single gel is likely to have hundreds of spots. In this paper, simulation method is used to generate reference gel and target gel not in the level of image but in the level of point pattern. The correspondence can be easily obtained by simulation method.

A reference gel with 500 spots is randomly generated with the minimum distance of 5 pixels and the image size of 512 by 512 pixels. A target gel is generated by distorting the reference gel with random values of x and y using the normal distribution. The same spot numbers are assigned both in the reference and in the target gel to conveniently check ground truth of the correspondence, meaning that one matched pair

with the same spot number is correctly matched. In this paper, we assume missing spots or outlier spots do not exist for simpler trial.

3.2 Experiment and result

The experiments are performed for two cases. One is for spot matching by the first neighbor spots with 5-NNG proposed in the literature [10]. The other experiment is for the second neighbor spots matching with 5-NNG proposed in this paper. The experiment results are summarized in the Table 1 and the Table 2.

The Table 1 shows only top 10 best matched pairs from the whole result by the methods of the first neighbor spots and the second neighbor spots, respectively, because it is too long to present all the result in this paper even though it is top 5% rankings. The labels 'R' and 'T' mean reference spot number and target spot number. The labels 'M', 'U', and 'H' mean the

number of matched pairs, the number of unmatched spots and the normalized Hausdorff distance in pixel, respectively.

In the first experiment of the Table 1, there are two falsely matched pairs displayed in the shaded areas such as $(R,T)=(358,47)$ and $(16,63)$. They are not right matches but they are matched with high similarity because they have similar topology by chance. It seems that it is a serious problem to have some false-positive results in the spot matching method by the first neighbor spots despite these are top 10 best matches. On the contrary, top 10 best pairs from spot matching method using the second neighbor spots have the same spot numbers as reference spot number and target spot number, which is interpreted as correctly matched.

The Table 2 shows best true-positive matches for top 10% and 20% in both cases of spot matching methods and it also shows true-positive matches for the whole result from both cases. In the column of true-positive pairs of top 10%, it

Table 1. Top 10 pairs from the first and the second neighbor spots

No. of experiment	Spot matching(5-NNG)									
	The first neighbor spots					The second neighbor spots				
	R	T	M	U	H	R	T	M	U	H
1	197	197	16	0	0.17	48	48	44	24	0.47
2	487	487	15	1	0.18	164	164	42	12	0.54
3	93	93	14	0	0.18	10	10	39	4	0.37
4	91	91	14	0	0.28	334	334	38	24	0.77
5	300	300	14	0	0.28	197	197	37	14	0.6
6	157	157	14	0	0.29	330	300	37	23	0.35
7	296	296	14	0	0.38	253	253	36	7	0.39
8	358	47	14	0	0.76	70	70	36	16	0.33
9	164	164	14	1	0.20	418	418	35	1	0.43
10	16	63	14	1	0.82	51	51	35	2	0.49

Table 2. Results of Spot matching using the first and the second neighbor spots

Measures	Spot matching method(5-NNG)			
	The first neighbor spots		The second neighbor spots	
		%		%
True positive pairs of top 10% best matches	20	40.0	50	100
True positive pairs of top 20% best matches	44	44.0	99	99.0
Total true positive pairs	164	12.8	330	66.0

has only 20(40%) of true positive matched pairs out of 50 pairs in the spot matching method by the first neighbor spots, meaning that it is quite inappropriate for the detection method of landmark matched pairs. It has 100% of true positive rate in the case of spot matching method by the second neighbor spots, meaning that it is a quite valid method to provide a large amount of correct landmark matched pairs. Spot matching method by the second neighbor spots seems to be very satisfactory even in the case of true positive pairs of top 20% best matches showing very high accuracy of 99%.

One or more matched pairs are required as initial values in a spot matching algorithm based on landmark matched pairs, which entirely depends on the algorithm itself. Researchers can get any number of reliable matched pairs by applying different criterions. Also, they can take top 15% or 20% of the whole matched pairs if the higher degree graphs such as 6-NNG, 7-NNG and 8-NNG are used to get more reliable matched pairs as initial input.

4 Conclusion

In this paper, a fully automated detection method of matched pairs for landmark spots is proposed. It uses topological patterns of the second neighbor spots to reduce occurrence of similar topological patterns in the case of densely populated gel image. The second neighbor spots help the proposed method to produce more accurate and robust matched pairs in the field of 2-DE spot matching with severe variation.

The results from the automated detection can be used in many previous spot matching methods which need matched pairs as initial input. The proposed method can detect as many matched pairs as needed by the algorithms if it takes more percentage of top ranked matched pairs instead of top 10%. The manual detection for landmark matched pairs can be replaced by the proposed method and expected to enhance the spot matching processes by removing the bottleneck.

Further study should be performed with a plenty of spot matching cases to improve the proposed method. And the study on the effects of graphs is required when various kinds of graphs and higher degree of k-NNG graphs are applied to the proposed method.

5 References

- [1] Y.-S. Hwang and J.-H. Lee, "Matching spots in Electrophoresis Images by Topology Preserving Relaxation (in Korean)," *The Korean Institute of Information Scientists and Engineers: Software and Application*, Vol. 39, No. 6, pp. 436-443, 2012.
- [2] J. L. Harry, M. R. Wilkins and B. R. Herbert, "Proteomics: Capacity versus Utility," *Electrophoresis*, Vol. 21, pp. 1071-1081, 2000.
- [3] P. H. O'Farrell, "High Resolution Two-Dimension Two-Dimensional Electrophoresis of Proteins," *Journal of Biological Chemistry*, Vol. 250, No. 10, pp. 4007-4021, 1975.
- [4] M. Daszykowski, E. Mosleth Faeregestad, H. Grove, H. Martens, B. Walczak "Matching 2D Gel Electrophoresis Images with Matlab Image Processing Toolbox," *Chemometrics and Intelligent Laboratory Systems*, Vol. 96, pp.188-195, 2009.
- [5] M. Rogers, J. Graham and R. P. Tonge, "2 Dimensional Electrophoresis Gel Registration Using Point Matching and Local Image-Based Refinement," *Proceedings of the 15th British Machine Vision Conference*, 2004.
- [6] A. Almansa, M. Gerschuni, A. Pardo and J. Preciozzi, "Processing of 2D Electrophoresis Gels", *1st International Workshop on Computer Vision and Applications for Developing Regions (ICCV)*, 2007.
- [7] Chang, S.H Cheng, F.H., Hsu, W.H., Wu, G.Z., "Fast Algorithm for Point Pattern-matching: Invariant to Translations, Rotations and Scale Changes", *PR20*, pp. 311-320, 1997.
- [8] J. Salmi, T. Aittokallio, J. Westerholm, M. Griese, A. Rosengren, T. A. Nyman, R. Lahesmaa, O. Nevalainen "Hierarchical Grid Transformation for Image Warping in the Analysis of Two Dimensional Electrophoresis Gels", *Proteomics*, Vol. 2, pp. 1504-1515, 2002.
- [9] Yun-Kyoo Ryoo, Chan-Myeong Han, Ja-Hyo Ku, Dae-Seong Jeoune, and Young-Woo Yoon, "Grassfire Spot Matching Algorithm in 2-DE," *International Journal of Bio-Science & Bio-Technology*, Vol. 5, No. 4, pp. 167-174, 2013.
- [10] Chan-Myeong Han, Dae-Seong Jeoune, Hwi-Won Kim, and Young-Woo Yoon, "A Spot Matching Algorithm using the Topology of Neighbor Spots in 2D-PAGE Images," *International Journal of Software Engineering and Its Applications*, Vol. 7, No. 5, pp. 87-97, 2013.

Object Pose Dataset using Discriminatively Trained Deformable Part Models

Jinho Kim¹, Yu Xiang², and Silvio Savarese³

¹Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea

²Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, USA

³Computer Science Department, Stanford University, Stanford, CA, USA

Abstract - Over the last couple of years computer vision has grown. While the old problem used to be object detection, we are now faced with the challenge of correctly estimating the pose of the objects. Thus in order to test algorithms for pose estimation it is important to use good datasets for training data. However the object datasets we have today are mainly for object detection. Therefore we do not have many sufficient datasets suitable for testing pose estimation algorithms. In this paper we use deformable part models and latent SVM to propose a dataset that we hope can become a good dataset for testing pose estimation algorithms.

Keywords: Deformable Part Models, latent SVM, Object Pose Estimation, ImageNet, PASCAL VOC

1 Introduction

The human ability to correctly classifying objects no matter what the pose of the object is something that cannot be replicated easily. As computer vision becomes more important, our interests are focused on how to correctly estimate the pose of objects. Thus in order to test algorithms of pose estimation it is important to use good datasets for training data. Savarese and Fei-Fei created a 3D object dataset that contained 10 categories and between 480 to 720 images per category [1]. The EPFL Car dataset has around 2300 images for 20 car instances [2]. While they continue azimuth, there is no variation to elevation or distances. Later PASCAL VOC datasets were created [3]. These datasets were intended for object detection not for pose estimation and thus had only four discrete viewpoints available, 'front', 'back', 'left', and 'right'. Our goal for this project was to create a large scale dataset for pose estimation that has a) thousands of images b) contains many categories of both indoor and outdoor objects and finally c) contains a large variation of viewpoints in terms of azimuth, elevation, and distance. In order to complete this dataset, we first start by amassing a large number of images of object categories from various image databases. We combined the images from the ImageNet database and PASCAL VOC images. However these databases were not meant for object

pose estimation thus we created an annotation tool based on 3D CAD models to compute the viewpoint corresponding to 2D and 3D objects. After collecting the images we made deformable part models which we used for training. In section 2.1 we will explain how we prepared our images for the dataset. In sections 2.2 and 2.3 we will briefly explain the concept of deformable part models and latent SVM respectively. Section 3 will deal with the results.

2 Methodology

2.1 Images

As stated before by just using the PASCAL VOC datasets presents us with some faults. Many of its images apart from the invariance in viewpoint present occlusion and truncation. This presents problems with trying to train the data as we want to training images to be as clean as possible. Thus when selecting images from ImageNet we attempted to select clean images with many more different viewpoints for training. We used an anchor annotator, created by Yu Xiang, to define anchor points on 3D CAD models. Annotators click on the anchor points in 2D images. After the annotator clicks on the anchor points the computer computes the viewpoint using 2D – 3D correspondence. Figure 1 shows the anchor annotator.

2.2 Deformable Part Models (DPM)

Once we have the images gathered we used models that are trained discriminatively so that they only require bounding boxes for the objects in the image. This leads to more efficient object detections. The training process returns a model that is a mixture of star models produced in the process. The testing process of DPM uses feature pyramids to detect features of the object contained within the image. From these results we test various threshold values to find which value gives the highest recall and precision. Here precision is the fraction of the reported bounding boxes that are correct detections while recall is the fraction of the objects found.

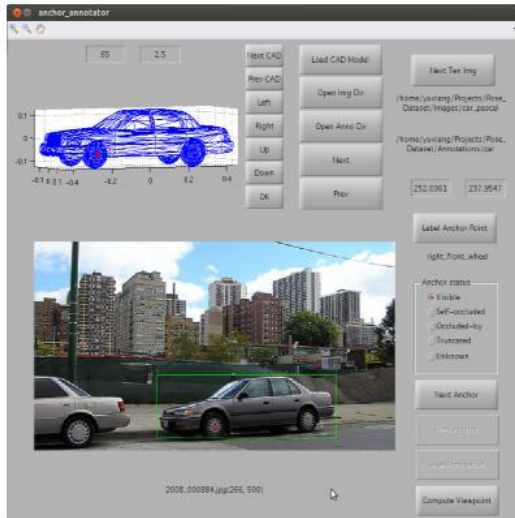


Fig. 1. The Anchor Annotator

The main approach is built on pictorial structures framework [4] [5]. Pictorial structures represent objects by a collection of parts arranged in a deformable configuration. Each part captures local appearance properties of an object while the deformable configuration is characterized by spring-like connections between certain pairs of parts.

The goal for DPM is to model objects using “visual grammars”. Grammar based models [6] [7] [8] generalize deformable part models by representing objects using variable hierarchical structures. Each part in a grammar based model can be defined directly or in terms of other parts. Also grammar based models allow for, and explicitly model, structural variations. These models provide a natural framework for sharing information and computation between different object classes.

The Dalal-Triggs detector [9] used a single filter on histogram of oriented gradients (HOG) features to represent an object category. The detector determines whether or not there is an instance of the target category at the given position and scale. The first innovation DPM uses involves enriching the Dalal-Triggs model using a star-structured part-based model defined by a “root” filter plus a set of parts filters and associated deformation models [10]. Figure 2 and Figure 3 each show a star model for the bicycle and person category, respectively, from the original PASCAL dataset.

2.2 Latent SVM

To train models using partially labeled data we use a latent variable formulation of MI-SVM [11] that is called latent SVM (LSVM). In a latent SVM each example x is scored by a function of the following form,

$$f_{\beta}(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z) \quad (1)$$

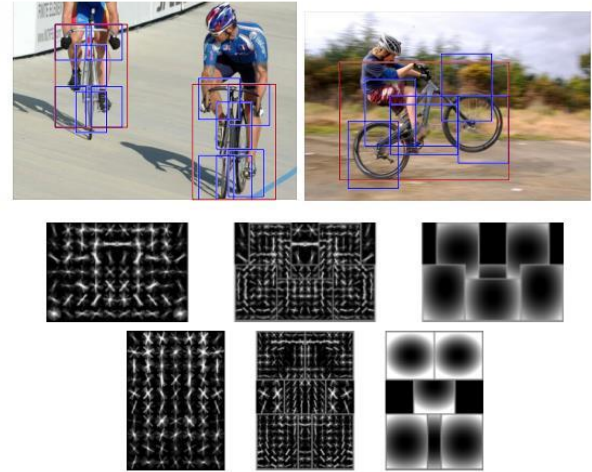


Fig. 2. Detections obtained with a 2 component bicycle model. These examples illustrate the importance of deformations mixture models. In this model the first component captures sideways views of bicycles while the second component captures frontal and near frontal views. The sideways component can deform to match a wheel. It should be noted that this is not from our dataset but from [10].

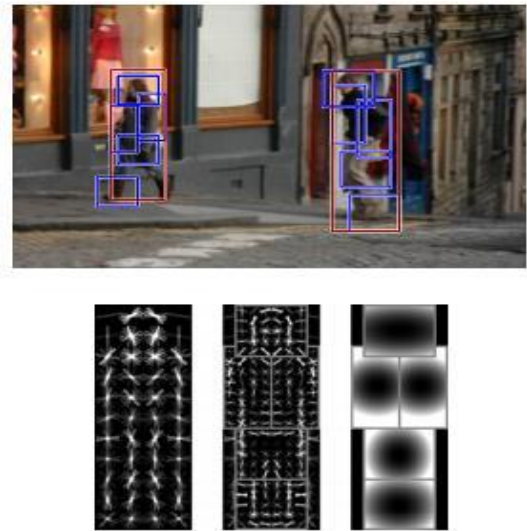


Fig. 3. Detections obtained with a single component person model. It should be noted that this is not from our dataset but from [10].

Here β is a vector of model parameters, z are latent values, and the function of (x, z) is a feature vector. In the case of one of our star models β is the concatenation of the root filter, the part filters, and deformation cost weights, z is a specification of the object configuration, and (x, z) is a concatenation of subwindows from a feature pyramid and part deformation features. It should be noted that equation (1) can handle very

general forms of latent information. To obtain high performance using discriminative training it is often important to use large training sets.

3 Results

We evaluated the performance of deformable part models on the dataset we constructed. It should be noted that this is still a trial run and we are still modifying our dataset to obtain more optimal results. Each dataset contains thousands of images of real world scenes. The datasets specify ground-truth bounding boxes for several object classes. At test time, the goal is to predict the bounding boxes of all objects of a given class in an image. Theoretically, a system will output a set of bounding boxes with corresponding scores, and we can threshold these scores at different points to obtain a precision-recall curve across all images in the test set. For a particular threshold the precision is the fraction of the reported bounding boxes that are correct detections, while recall is the fraction of the objects found.

A predicted bounding box is considered correct if it overlaps more than 50% with a ground-truth bounding box, otherwise the bounding box is considered a false positive detection. Multiple detections are penalized. If a system predicts several bounding boxes that overlap with a single ground-truth bounding box, only one prediction is considered correct, the others are considered false positives. One scores a system by the average precision of its precision-recall curve across a testset. Figure 4 shows a sample precision-recall curve for the category of cars.

In some categories our false detections are often due to confusion among classes, such as between car and bus. In other categories false detections are often due to the relatively strict bounding box criteria. The two false positives shown for the person category are due to insufficient overlap with the ground truth bounding box.

4 Conclusions

Deformable part models have been proven on various datasets. [10] Therefore we must retest it after finalizing our dataset. As of now we have determined that our dataset needs further evaluation and improvements. We realize that it is not sufficient for us to use this dataset for experimentation on object pose detection. Therefore we must add more images and discard unfit images which we are currently doing at the

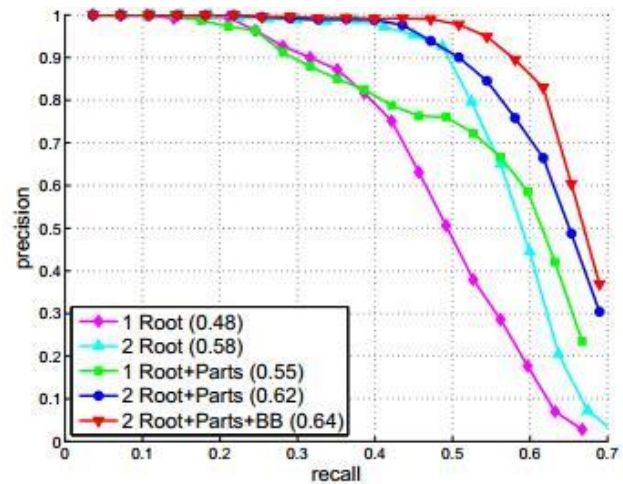


Fig. 4. Precision/Recall curve for models trained on the car category. We show results for 1 and 2 component models with and without parts, and a 2 component model with parts and bounding box prediction. In parenthesis we show the average precision score for each model. It should be noted that this is not from our dataset but from [10].

moment. It is also a good idea to add more categories or to divide the pre-existing categories into many separate subcategories. As of now we are exploring the idea of dividing the categories into separate subcategories. While it is important to have a big dataset for accurate testing, it is also important to select the best images. Since the deformable part model was to test our dataset we can rely on the results to show where we must improve our dataset.

References

- [1] Savarese, S., Li, Fei-Fei. "3D Generic Object Categorization, Localization and Pose Estimation"; IEEE International Conference in Computer Vision, IEEE Press, New York, pp. 1-8, 2007.
- [2] Ozuysal, M., Lepetit, V., Fua, P. "Pose Estimation for Category Specific Multiview Object Localization"; IEEE Conference on CVPR, IEEE Press, New York, pp. 778-785. 2009.
- [3] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. "The PASCAL Visual Object Classes (VOC) Challenge"; International Journal of Computer Vision. Vol. 88, pp. 303-338, 2010.
- [4] Felzenszwalb, P. and Huttenlocher, D. "Pictorial structures for object recognition"; International Journal of Computer Vision. Vol. 61, pp. 55-79, 2005.

- [5] Fischler, M. and Elschlager, R. "The representation and matching of pictorial structures"; IEEE Transactions on Computer. Vol. 22, pp. 67-92, 1973.
- [6] Felzenszwalb, P. and McAllester, D. "The generalized A* architecture"; Journal of Artificial Intelligence Research. Vol. 29, pp. 153–190, 2007.
- [7] Jin, Y., Geman, S. "Context and hierarchy in a probabilistic image model"; IEEE International Conference in Computer Vision, IEEE Press, New York, pp. 2145-2152, 2006.
- [8] Zhu, S., Mumford, D. "A stochastic grammar of images"; Foundations and Trends in Computer Graphics and Vision. 2, pp. 259–362, 2007.
- [9] Dalal, N., Triggs, B. "Histograms of oriented gradients for human detection"; IEEE Conference on CVPR, pp. 886-893, 2005.
- [10] Felzenszwalb, P., Girchick, R., McAllester, D., Ramanan, D. "Object Detection with Discriminatively Trained Part Based Models"; IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 32, pp. 1627-1645 , 2010.
- [11] Andrews, S., Tsochantaridis, I., Hofmann, T. "Support vector machines for multiple-instance learning"; Advances in Neural Information Processing Systems. Vol. 15, pp. 561-568 , 2003.

HDR Image Appearance Mapping Using Dual Exposed LDR Images

Hyuk-Ju Kwon[†], Sung-Hak Lee[†], Geun-Young Lee[†] and Kyu-Ik Sohng[†]

[†]School of Electronics Engineering, Kyungpook National University
1370 Sankyug-Dong, Buk-Gu, Daegu 702-701, Korea
E-mail: shak2@ee.knu.ac.kr

Abstract - Display devices are difficult to express the high contrast scene because the dynamic range of device is very low than the real scene or human visual system (HSV). High dynamic range (HDR) tone mapping algorithms are developed to create a realistic image. An HDR blending algorithm is the combination method for multiple exposed low dynamic range (LDR) images and constructs an HDR radiance map. Existing HDR blending algorithms show color distortion when using only two LDR images. The reproduction of camera response function (CRF) is hard to be estimated by few images because of the lack of dynamic range information. iCAM is the HDR tone mapping algorithm based on the color appearance model. However, iCAM has hue and chroma distortion after tone compression step. This paper proposes an HDR blending and a tone mapping model to improve dual image-based HDR image toning.

Keywords: HDR, Radiance map, Tone mapping, iCAM, CIECAM02, CRF

1 Introduction

Digital cameras and display devices which have low dynamic range (LDR) are limited to capture or display for a real scene. High dynamic range (HDR) tone mapping methods are developed to display the dynamic range of a real scene efficiently. HDR blending enhances the image dynamic range by combining multiple exposure images of LDR [1,2]. Additionally HDR tone mapping is supposed to reduce or transform the dynamic range of images for showing HDR images on LDR devices.

Debevec et al. [3] proposed the recovered HDR radiance map from photographs. This algorithm recovers the camera response function (CRF) from a series of different exposure images and the radiance map is recovered from these CRF. But, this algorithm needs a lot of pictures to find the accurate CRFs. Overall color distortion occurs as a result of the inaccurate CRFs when the radiance maps are computed using only two pictures. Image color appearance model (iCAM) as one of HDR tone mapping algorithms based on human visual system (HVS) has chromatic adaptation and tone mapping steps [4,5,6]. Chromatic adaptation step predicts the illuminant of an image and changes colors to corresponding colors according to a viewing condition. Tone mapping step

reduces the dynamic range of an image in accordance with the cone responses of HVS. However, tone mapping step causes the unpredictable change of hue and chroma as well as dynamic range. This causes color appearance distortion. Therefore, there is a need to improve it.

In this paper, we propose a HDR blending algorithm and a tone mapping model for the improvement of hue and chroma distortion after tone mapping. To reduce the distortion of CRFs, the new HDR blending algorithm includes a virtual middle exposure image and variable weighting functions for each exposure image. The proposed tone mapping has parallel processing for chromatic adaptation and tone mapping to reduce interference effect of each step. Color appearance mapping based on CIECAM02 [7] is applied after tone mapping step to preserve the hue and chroma in chromatic adaptation step.

2 Related Works

2.1 Radiance Image Mapping

Debevec et al. algorithm constructs HDR radiance maps by combining images with different exposures for a same scene. Camera response function (CRF) is needed in order to calculate the radiance maps. Figure 1 shows the block diagram of the Debevec's algorithm. Equation 1 represents the relationship between CRF and irradiance E .

$$g(Z_{ij}) = \ln E_i + \ln \Delta t_j \quad (1)$$

where i is a spatial location and j is an exposure duration. g is CRF, Z is pixel brightness value, E is irradiance, and Δt is exposure time. In Eq. 1, the function g and the irradiance E are corresponding to the over determined system. One approach to find solution is the least square method. The least square method for g is represented as follows.

$$O = \sum_i^N \sum_j^P \left\{ \omega(z_{ij}) [g(Z_{ij}) - \ln E_i - \ln \Delta t_j]^2 \right\} \quad (2)$$

$$+ \lambda \sum_{z=Z_{\min}+1}^{Z_{\max}+1} [\omega(z) g''(z)]^2 \quad (3)$$

$$g''(z) = g(z-1) - 2g(z) + g(z+1)$$

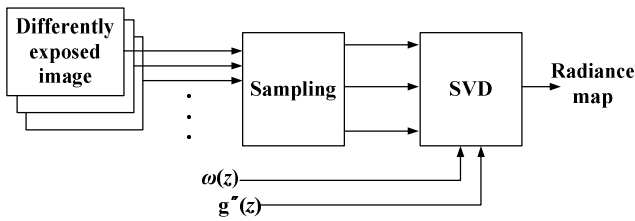


Fig. 1. Block diagram of the Debevec's algorithm.

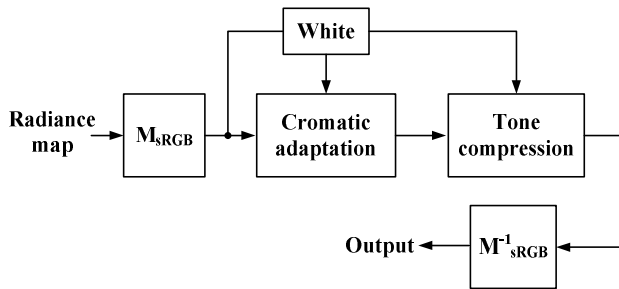


Fig. 2. Block diagram of iCAM.

The second term of Eq. 2 is derived to ensure that the function g is smooth. $\omega(z)$ is a weighting function to emphasize the smoothness and fitting terms toward the middle of the function g . λ is smoothness value for data fitting term. Equation 2 is derived using the singular value decomposition (SVD) method. The HDR radiance map is constructed as follow.

$$\ln E_i = \sum_{j=1}^p \omega(Z_{ij})(g(Z_{ij}) - \ln \Delta t_j) / \sum_{j=1}^p \omega(Z_{ij}) \quad (4)$$

2.2 Image Color Appearance Model

iCAM is an image rendering model based on CIECAM02. CIECAM02 is a color appearance model that estimates brightness, lightness, colorfulness, chroma, saturation, and hue. iCAM and iCAM06 are representative HDR tone mapping algorithms and these algorithms have the chromatic adaptation and tone compression that consider the human visual properties according to luminance level. Figure 2 represents the brief block diagram of iCAM. The chromatic adaptation converts the tri-stimulus values to adapted tri-stimulus values. It applies the human visual adaptation in accordance with the variation of viewing conditions. The tone compression processing reduces the dynamic range of the radiance map to display LDR devices. The tone compression functions adopt the photoreceptor response curves. Input tri-stimulus values take through chromatic adaptation and tone compression based on the luminance level of each pixel after estimating illumination components of the image. A white image from the Gaussian low pass filter is used to viewing condition data of each processing.

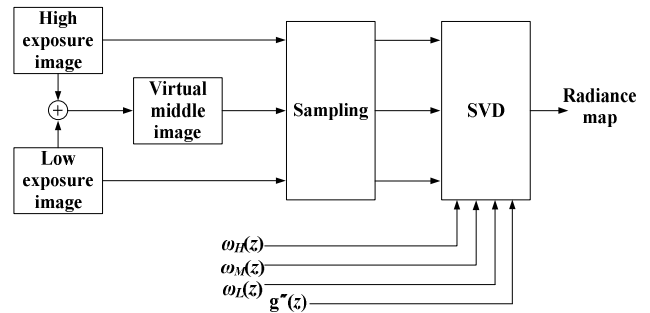


Fig. 3. Block diagram of the modified radiance mapping algorithm.

3 Proposed Dual Image Blending and Tone Processing

A new HDR blending algorithm makes a virtual middle exposure image from the average of high and low exposure images. This middle exposure image provides the middle dynamic range information for the middle exposure data of the SVD. Additionally the variable weighting functions with high, middle, and low positions are applied for each exposure image that provides different visual information. The variable weighting functions complement the dynamic range of each image. As a result, the error of CRFs and color distortion are reduced. Figure 3 shows the modified radiance mapping algorithms. Equations (5)-(7) provides the weighting functions.

$$\omega_H(z) = \begin{cases} z - Z_{\min} & \text{for } z \leq \frac{1}{2}(Z_{\min} + Z_{\max}) \\ 1 & \text{for } z > \frac{1}{2}(Z_{\min} + Z_{\max}) \end{cases} \quad (5)$$

$$\omega_M(z) = \begin{cases} z - Z_{\min} & \text{for } z \leq \frac{1}{2}(Z_{\min} + Z_{\max}) \\ Z_{\max} - z & \text{for } z > \frac{1}{2}(Z_{\min} + Z_{\max}) \end{cases} \quad (6)$$

$$\omega_L(z) = \begin{cases} 1 & \text{for } z \leq \frac{1}{2}(Z_{\min} + Z_{\max}) \\ Z_{\max} - z & \text{for } z > \frac{1}{2}(Z_{\min} + Z_{\max}) \end{cases} \quad (7)$$

where z is the pixel brightness value. Z_{\min} and Z_{\max} are the minimum and maximum pixel values.

In iCAM, the dynamic range variation by tone mapping influences the hue and chroma then it changes chromatic adaptation points and color appearance of input images because iCAM has high interference effect between chromatic adaptation and tone mapping. To reduce the change of hue and chroma, chromatic adaptation and tone mapping are

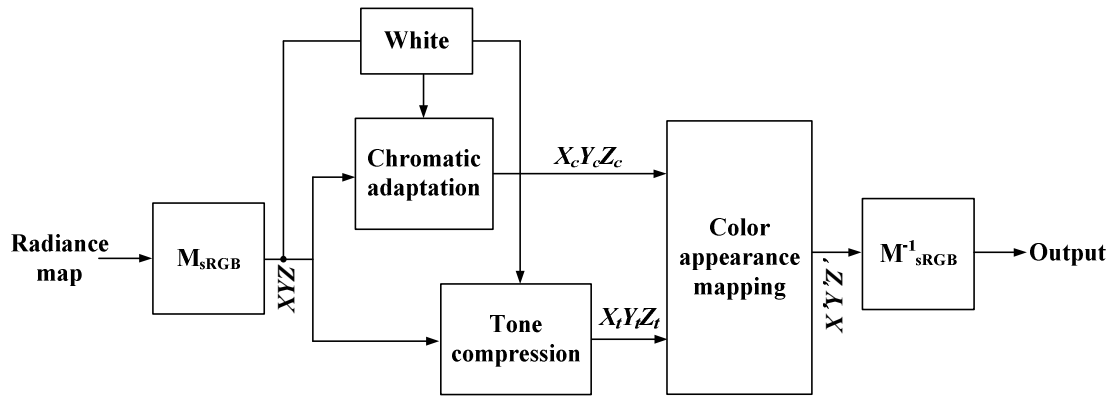


Fig. 4. Block diagram of the proposed HDR tone mapping.

processed separately and color appearance mapping based on CIECAM02 is applied after the parallel processing. The block diagram of a proposed method is shown in Fig. 4. Input image is a radiance map consisting of two different exposure images and one virtual exposure image. The radiance map is converted to CIE tri-stimulus using the transform matrix M_{sRGB} . Chromatic adaptation uses a RGB cone response converted from CIE tri-stimulus using the transform matrix M_{CAT02} . White represents an adaptation luminance level to calculate the corresponding colors of each pixel. Tone mapping uses a $R'G'B'$ cone response converted from CIE tri-stimulus using the transform matrix M_{HPE} .

The dynamic range of input image is reduced by tone mapped cone responses. Color appearance mapping step uses the CIECAM02 model to restore the hue and chroma after tone mapping. Color appearance of objects can be calculated from the results of chromatic adaptation and tone compression. Lightness (J) comes from tone mapping results, hue (h) and chroma (C) come from chromatic adaptation results. We calculate tone mapped lightness (J_t) of $X_t Y_t Z_t$, and chromatic adapted hue (h_c) and chroma (C_c) of $X_c Y_c Z_c$. Output image is converted from J_t , h_c , and C_c using inverse CIECAM02 (f^{-1}). For remapped RGB responses, an inverse function is derived from J_t , h_c , and C_c as follows.

$$R'_a = f_R^{-1}(J_t, h_c, C_c) \quad (8)$$

$$G'_a = f_G^{-1}(J_t, h_c, C_c) \quad (9)$$

$$B'_a = f_B^{-1}(J_t, h_c, C_c) \quad (10)$$

The CIE tri-stimulus $X'Y'Z'$ is obtained by the inverse matrix of M_{CAT02} from $R'G'B'$. Finally, output RGB image is converted from the CIE tri-stimulus using the transform matrix M_{HPE} .

4 Simulation Results

We compare the proposed method to a conventional method in Fig. 5 and 6. Figure 5 shows the HDR blending algorithms. Fig 5(a) is input images which have different exposure for a same scene, Fig. 5(b) is a result by Debevec's algorithm and Fig. 5(c) is a result by the proposed algorithm. Figures show the CRFs of each algorithm. In Fig. 5(b), hue shift is represented at lower and higher levels. Because the red channel of CRFs isn't uniformly distributed in red circle areas. The proposed radiance mapping algorithm shows uniform distribution and the hue shift is improved in Fig. 5(c).

Figure 6 shows the results of iCAM06 and the proposed tone mapping model. Above the HDR blending algorithm is used to input image of each model. In Fig. 6(a), the rendered image of iCAM06 shows the hue shift for an entire image even if input radiance image be improved. The detail and contrast of iCAM06 are lower than the proposed tone mapping model. In Fig. 6(b), the proposed method shows an enhanced image in terms of the local contrast, details, and chroma, colors of building and clothes are similar to those of input images. Detail and contrast are decreased in bright or dim areas of the image.

5 Conclusions

The HDR blending algorithm can reproduce the radiance image similar to the dynamic range of human vision using multi exposure images captured by digital cameras. Then HDR images can be represented on an LDR display device using tone mapping algorithm. Tone mapping can reduce the dynamic range of HDR images. But, existing HDR algorithm needs many exposure images to calculate accurate CRFs and tone mapping method cause the distortion of hue and chroma. This paper proposes the novel HDR blending and tone mapping method to overcome these problems.

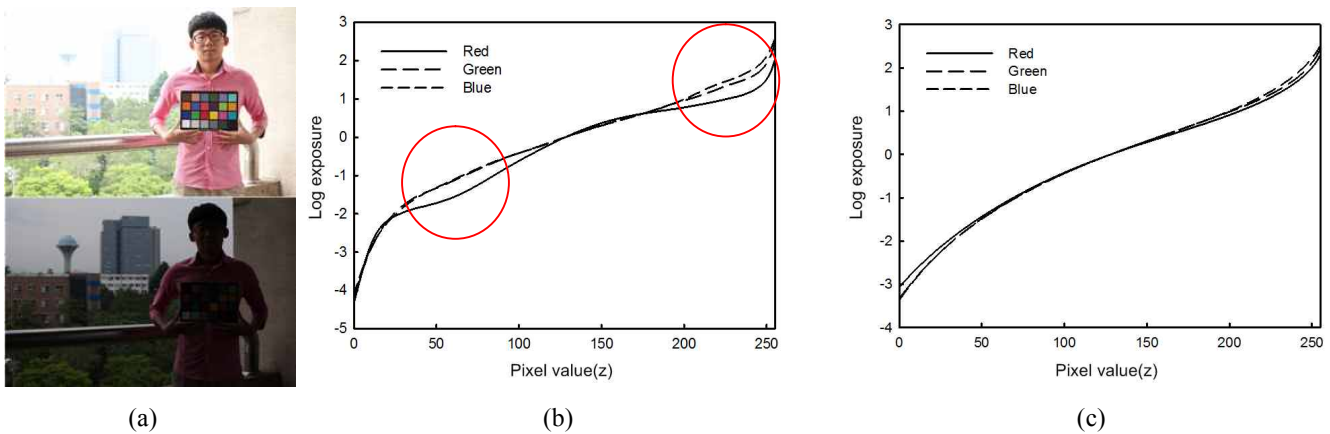


Fig. 5. Input images and HDR blending. (a) Input images, (b) Debevec's algorithm, and (c) the proposed blending.

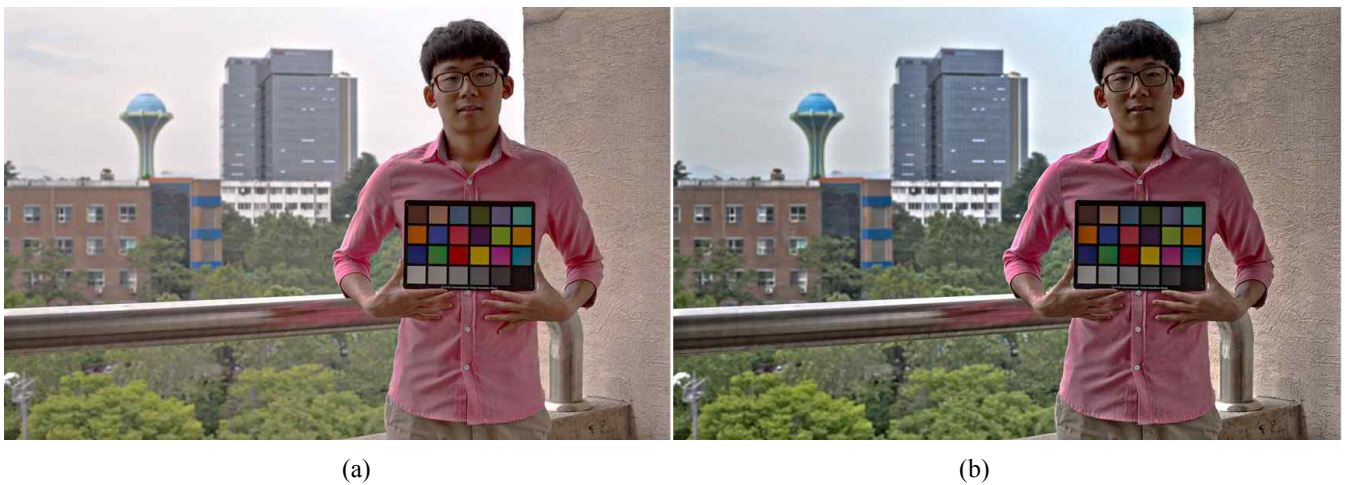


Fig. 6. HDR rendered images. (a) Debevec's blending and iCAM06 and (b) the proposed blending and tone mapping.

The proposed HDR blending uses two exposed images and makes the virtual middle exposure image for the supplementary middle dynamic range data then the variable weighting function is applied to a SVD method. The result of HDR radiance map shows the uniform distribution curve of CRFs and the enhancement of color distortion. The proposed tone mapping method consists of the parallel processing steps of chromatic adaptation and tone mapping to reduce the relevance of two steps. Color appearance mapping is applied to correct corresponding colors. In simulation results, the proposed method shows obvious improvement in local contrast, detail, and color preservation aspects.

6 References

[1] G. Johnson, M. Fairchild, "Rendering HDR Images," in: *IS&T/SID 11th Color Imaging Conference, Scottsdale*, pp. 36-41, 2003.

[2] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, Morgan Kaufmann, 2005.

[3] P. Debevec and J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," *SIGGRAPH '97*, pp. 369-378, 1997.

[4] M. Fairchild, *Color Appearance Models 2nd Ed.*, John Wiley & Sons Press, 2005.

[5] M. Fairchild and G. Johnson, "iCAM Framework for Image Appearance, Differences, and Quality," *J. Electron. Imaging.*, vol. 13(1), pp. 126-138, 2004.

[6] J. Kuang, G. Johnson, and M. Fairchild, "iCAM06: a refined image appearance model for HDR image rendering", *Journal of Visual Communication and Image Representation*, vol. 18, no. 5, pp. 406-414, 2007.

[7] N. Moroney, M. Fairchild, R. Hunt, C. Luo, and T. Newman, "The CIECAM02 Color Appearance Model," in: *IS&T/SID 10th Color Imaging Conference, Scottsdale*, pp. 23-27, 2002.

SESSION

OBJECT DETECTION, RECOGNITION, AND CLASSIFICATION + OCR, AND RELATED ISSUES

Chair(s)

TBA

OBJECT DETECTION AND MATCHING USING PROPOSED SIGNATURE AND SURF

Hany A. Elsalamony*

Mathematics Department, Faculty of Science, Helwan University, Cairo, Egypt

h_salamony@yahoo.com

Abstract - Most of algorithms of object detection and classifications is only locating region in the image, whether it is within a template-sliding mask or interested region blobs. However, such regions may be ambiguous, especially when the object of interest is very small, unclear, or anything else. This paper presents proposed algorithm for automatic object detection and matching based on its own proposed signature using morphological segmentation tools. Moreover, the algorithm tries to match the objects; neither among object's blobs nor among regions of interest; but among the constructed proposed objects' signatures. In the matching process, SURF method has been presented to make a comparison on the experimental results. The performance has been tested 120 from a wide variety of unlike objects, it has been achieved 100% in the case of constructing object signatures, also it has been achieved 96% of right matching whereas SURF has achieved 85% for all experimental objects.

Keywords: Object Detection and Matching; Signature; SURF; Segmentation.

1 Introduction

The object detection plays an important role in the area of computer vision research. Nowadays, many of its applications require the locations of objects in images. In fact, there are two closely related definitions, object presence detection, and object localization. The determination of one or more of an object class are present (at any location or scale) in an image, that means of an object's presence detection or image classification, and can be suitable for image retrieval based on an object [6]. While the object localization means finding the object location and scale an image.

Many of the object detection algorithms are following the model of detection by parts; that introduced by Fischler and Elschlager [18]. They are used the object structural modeling and reliable part detectors' methods. The basic idea behind this model is to identifying that the individual parts of an object detector are easier to build than that for the full object [8], [14]. Actually, these methods of object detection are depending on sliding a window or template mask through the image to classify each object falls in the a local window as a background

or target [5], [13]. In fact, this approach has been successfully used to detect rigid objects such as cars and faces, and has even been applied to articulated objects such as pedestrians [4], [11], [19].

Later, a frequency model proposed, which is dependent on a moving background containing repetitive structures. The authors are considered special temporal neighborhoods of the pixels, which they have applied local Fourier transform in the scene [3]. The feature vectors, which generated are used to build a background model. However, they are applied their model for moving object and backgrounds, on both synthetic and real image sequences [12].

On the other hand, one popular approach is depending on extracting the local interest points through the image, and then classifies the regions, which contained these points, instead of looking at all possible sub windows as the previous [20]. The greatest common divisor of the above approaches is that they can fail when the regional image information is insufficient (target is very small or unclear), and this is considered as a weakness of them [7].

In this way, the image matching based on features is depending on analyze the extracted features and find the corresponding relationship between them [20]. The image matching is not accurate enough because the images are often noisy, in different illuminations and scales. Recently, extracted features are widely applied in the field of object matching. In 1999, the Scale Invariant Feature Transform (SIFT) presented by Lowe, when a robust descriptor and Difference-of-Gaussians (DoG) detector had been used [15], [16]. It is interesting to note that the advantages of SIFT; that it is applied on invariant rotation or image scale, is about its computation which is very hard to calculate and take a time because it needs to extract 128 dimensional descriptors to work. This problem was solved in 2008 by Bay, who proposed SURF, it is modeled by 64 dimensions. The experiments of SURF have assumed the integrated images to compute a rough approximation of the Hessian matrix, and this is tends to faster than SIFT [9], [10]. In 2009, Lue and Oubong compared SIFT and SURF, they are pointed out that SURF is better in performance, but it is not efficient in rotation changes [17]. In fact, the effective power of SURF has been reduced because the ignoring of features in geometric relationships [17], [21].

*Computer Science Department, Arts & Science College, Salman Bin Abdul Aziz University, Saudi Arabia.

This paper presents proposed algorithm depending on the object geometrical shape, and relationship between outer points of the objects' contours. It is divided into two parts; one is constructing own signature for any object in an image. Second part is matching operation among all object shapes' signatures to get exactly which they are describe. In addition, four steps have to process through these parts; constructing signatures for all objects in an image and saving them as data in the system. Secondly, constructing signatures for all test input objects. The comparisons between inputs and saved signatures, which they have determined before, have operated using statistical methods in the third step. Finally, these signatures have used to detect and define the objects in image.

In fact, the proposed approach introduces an idea to detect object dependent on its outer shape by constructing own signature which let the object to be free from its constrains such as rotation, size, its position in the image. This proposed idea may will used in many object detection fields like identifying the kinds of plants based on their shapes, distinguishing between kinds of fruits, and so on. The next section is discussing SURF method because it is the most important one in the object detection and matching ways.

The rest of this paper is organized as follows. Section 2 is introducing an overview on SURF. A brief overview on image segmentation has in section 3. In section 4, the proposed algorithm of detection and matching has been illustrated. The algorithm's experimental results are showing in Section 5, and the conclusion in Section 6.

2 Overview on SURF Method

The initial mention of SURF (Speeded Up to Robust Features) was by H. Bay in 2006. It has four major stages: Hessian matrix, localization of these points, orientation assignment, and descriptor, which is depended on Haar wavelet responses sum [10]. In the first, Hessian matrix, which is based on detection in scale space of interested points. Additionally, the determinant of Hessian matrix is used as a preference to look for local maximum value and the detection of SURF interested point is based on theory of scale space. Equation (1) illustrates in details the components of Hessian matrix. In this equation, there is a point $X=(x, y)$ in an image I , the Hessian matrix $H(X, \sigma)$ in X at scale σ is defined as follows:

$$H(X, \sigma) = \begin{pmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{pmatrix} \quad (1)$$

where $L_{xx}(X, \sigma)$ represents the convolution of the Gaussian second order partial derivative. $\frac{\delta^2 g(\sigma)}{\delta x^2}$ with the image I in a point X , and similarly for $L_{xy}(X, \sigma)$ with $\frac{\delta^2 g(\sigma)}{\delta x \delta y}$ and $L_{yy}(X, \sigma)$ by $\frac{\delta^2 g(\sigma)}{\delta y^2}$.

To speed up the convolution, 9×9 box filter is utilized to approximate integral image and the second-order Gaussian

partial derivatives with $\sigma = 1.2$, [10]. The symbols D_{xx} , D_{xy} , and D_{yy} , are denoting the convolution results approximations. The determinant of Hessian matrix is:

$$|H_{approx}| = D_{xx}D_{yy} - (wD_{xy})^2 \quad (2)$$

where w is recommended as 0.9, that is the relative weight of the filter responses [9], [10]. The step after is dividing the image into many regions, each one contains different scale image templates.

The second stage is the interested point localization. First step in this stage is setting a threshold to the detected Hessian matrix of extreme points. Second step, to obtain these points a non-maximum suppression in a $3 \times 3 \times 3$ neighborhoods are applied. The basis of selecting a feature point are that only the point with value bigger than the neighboring 26 points' value is chosen as a feature point [10].

The third stage is the orientation assignment, that is starting by calculate the Haar wavelet (Haar side: $4s$, where s is the scale at which the interest point was detected) responses in x and y direction within a circular neighborhood of radius $6s$ around the interest point. The responses have been centered at the interested point and weighted with Gaussian ($2s$). At that moment, the sum is calculated for all responses within sliding orientation window of size $\pi/3$ to estimate the leading orientation, then determining the sum of horizontal and vertical responses within the window. A local orientation vector is produced with the two collected responses, such that the longest vector over all windows defines the orientation of the interest point.

Last stage in SURF is the descriptor based on sum of Haar wavelet responses. For the extraction of the descriptor, the first step consists of constructing a square template region (the size is $20s$) oriented along the selected orientation and centered on the interested point. The region splits up regularly into smaller 4×4 square sub-regions. For each sub-region, Haar wavelet responses are computed at 5×5 regularly spaced sample points. Simply, the Haar wavelet response in horizontal direction is denoted by d_x , also, the Haar wavelet response in vertical direction by d_y . The responses d_x and d_y are first weighted with a Gaussian ($\sigma = 3.3s$) centered at the interested point. Moreover, the responses $|d_x|$ and $|d_y|$ are extracted to bring in information about the polarity of changes in the intensity. Hence, the structure of each sub-region has four-dimensional descriptor vector V_{sub} :

$$V_{sub} = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|) \quad (3)$$

From the previous, by multiplying all 4×4 sub-regions results in a descriptor vector of length are $4 \times (4 \times 4) = 64$ [9], [10]. Additionally, to judge whether the two feature points of images are matched or not, the distance of the characteristic vector between two feature points is calculated. Finally, It is interesting to note that SURF ignores the geometric

relationship between the features, which is very important characteristic of many objects in the image. For that reason, this paper presents proposed algorithm to detect and matching objects based on own constructed signatures. The next section introduces a brief overview on image segmentation.

3 Overview on Image Segmentation

The segmentation idea is splitting an image into many various regions containing every pixel with similar characteristics such that; texture information, motion, color, whereas the detection stage has to choose relevant regions and assign objects for further processing [1], [8]. In addition, these regions should strongly related to the detected objects or features of interest to be meaningful and useful for image analysis and interpretation. Actually, the transformation from grey scale or color image in a low-level image into one or more other images in a high-level image, which is depending on features, objects, and scenes, represents the first step in significant segmentation. Generally, the accurate partitioning of an image is the main challenging problem in image analysis, and the success of it depends on consistency of segmentation [20].

On the other hand, segmentation techniques are divided into either contextual or non-contextual. The non-contextual techniques do not care about account of special relations between features in an image and group pixels together based on some global attribute, e.g. gray level or color. However, contextual techniques mainly exploiting these relations, e.g. pixels with similar gray levels and close spatial locations grouped with each other [6], [8]. Actually, the proposed algorithm is trying to exploit the segmentation contextual techniques in object detection and classification. Next section illustrates in details the idea for the suggested algorithm.

4 The Proposed Algorithm

The proposed object detection and matching framework is dividing into three parts. They are consisting of segmentation, construction of objects' signatures in image and matching them to classify the object based on its signature [7]. The segmentation process represents the main stone in this algorithm, which is giving initial hypotheses of object positions, scales and supporting based on matching. These hypotheses are then refined through the object signature classifier, to obtain final detection and signatures matching results. Fig.1 describes all steps of the proposed algorithm.

This Figure starts by an example of original RGB image with all different objects, many morphological functions and filters (edge detection, erosion, dilation, determination the number of objects, watershed segmentation ...) are applied to enhance the work of this image. The areas, centroids, orientations, eccentricities, convex areas for every object can easily be determined. Moreover, the boundary points (x_{ij}, y_{ij}) for each object is calculated individually, where i represents the number of objects and j is the number of boundary points related to an object, these boundary points and the objects' previous

informations are saved to start construction of own proposed signature for every object based on all these informations. The relation among all these information and Euclidian distance from objects' centroids (x_{ic}, y_{ic}) is plotted and saved as an individual signature for each object, that is shown in the Fig.1 by one object. These signatures for all objects are saved and waiting for matching with any input object's signature, as in the experimental results section. Moreover, the contour is drawn around all objects and tracing the exterior boundaries of them.

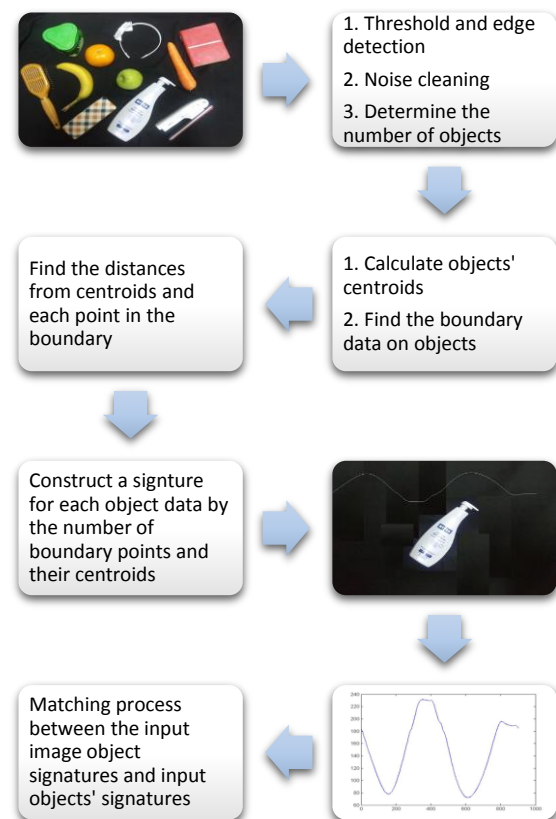


Fig.1. The proposed algorithm steps.

The third part of this proposed algorithm after segmentation and constructing signatures is the matching process between input and saved objects' signatures as in the above. Two different ways in matching process have been used to make a comparison between them in accuracy and activity; one is the using of statistical measures related to proposed signatures, and the second is using SURF.

Additionally, all shown steps in Fig.1 are also applied on the input object to construct its signature. Actually, the matching process is depending on statistical measures on both types of objects' signatures (saved ones and input). Firstly, as shown above in Fig.1, not all objects in the example image are in the same size, orientation, or even shape, and afterwards, their data should not be equal in length or characteristics. For that reason and more in checking accuracy, some pre-processes of matching have been carried out; one is sorting all the data of

all signatures, and then computing the variability in the data set by calculating the average of the absolute deviations of data points from their mean. The equation for average deviation is:

$$AVD = \frac{1}{n_i} \sum_{j=1}^{n_i} |x_{ij} - \bar{x}_i| \quad (4)$$

For all i 's are the number of objects in the image and j 's are the number of object's signature data, x_{ij} represents the number of signature's data points, \bar{x}_i is their mean, and n_i is the number of signature data rows. Secondly, the results of Equation (4) have been applied on all input and saved objects to make a comparison between them for get the exact matching by least error. Equation (5) introduces a method for calculate differences between the results of Equation (4).

$$DIF = ABS(ADV_{Saved} - ADV_{Input}) \quad (5)$$

The components of Equation (5) are the absolute value of the difference between the two results of Equation (4) related to saved and input object signature. The decision of matching based on the least value of DIF, which is given the exact matched object. Next section shows the experimental results of the proposed algorithm.

5 Experimental Results

The experimental results are divided into two parts; one is representing the objects and their signatures in images, and the second is showing the results of matching and comparing between proposed algorithm and SURF methods. Fig.2 presents a sample of experimental clear and unclear images, which contain some standard geometrical shapes in (a), some kinds of objects varying in shapes, and luminance intensity in (b), (c), (d), (e) and (f). Sequentially, the signatures have been constructed for the most distinct objects mentioned in Fig.2.

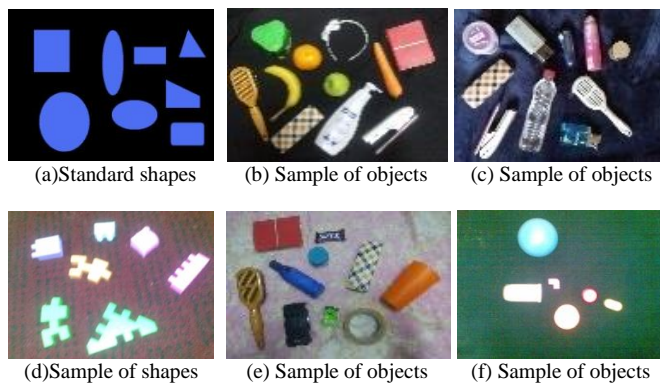


Fig.2. Images of famous regular shapes in (a) and from (b) to (f) other types of different objects

In Fig.3, all objects are individually be defined, detected, and matched by its signatures in the proposed algorithm. It is interesting to note that the similarity is cleared in signatures for the stand ellipse (its major axis parallel to the y-axis) and

horizontal one (its major axis parallel to the x-axis) because it is the same shape but different position. Evidently, the square shape has four identical peaks in its signature because the equality of its sides; furthermore, the circle's signature is one-line parallel to the x-axis and far away its radius length. In the same way, many objects' signatures are nearest each to others, for example, the object (1, 2) (i.e. in the row 1 and column 2 in Fig.3) is closer in signatures with objects in cells (9, 2), (9, 4). In the same context, signatures of objects (1, 3) and (5, 4) are seemed to correspond to some. These last cases are happened because the shape's nature of the original objects not because mistakes or big errors in the proposed algorithm.

Actually, the proposed algorithm has been applied on about 120 different shapes, positions, orientations, and intensity luminance of objects in RGB images. Furthermore, signatures determination for all objects has achieved 100% without any errors (in data, or wrong signature construction) for all objects.

		1	2	3	4	5	6
1	Objects						
2	Signatures						
3	Objects						
4	Signatures						
5	Objects						
6	Signatures						
7	Objects						
8	Signatures						
9	Objects						
10	Signatures						
11	Objects						
12	Signatures						

Fig.3. Different objects and their signatures

Second part of the proposed algorithm is matching of input and saved signatures. Table.1 presents the matching process that is depending on Equations (4), and (5); the decision in this process is based on the least value in Equation (5). Obviously, Table.1 shows all input objects images in first row. Regularly, the first column represents all positions (1 to 11) of objects in their main image if are scanned from left to right. Sequentially, that Table consists of x rows and y columns, which are contain a set of values, represent the least value of errors calculated by Equation (5). For example, in the cell (1, 1), the proposed algorithm is selected a least value (0.1097) in row (1), which indicates to the first object in the image. Clearly, this value indicates to the exact object position selected in the main image of Fig.2. (b). In this case, the input object has completely different in its position and orientation; however, the proposed matching algorithm is overcome that and succeeded. In fact, all other objects have been matched by the same way and have achieved 100% for that image.

Table.1. the matching process of objects' signatures



















											
1	0.1097	23.32694	5.452305	3.462439	27.57506	14.90506	12.47566	25.51032	1.590916	5.577318	20.90324
2	23.23969	0.022453	17.89709	19.88695	4.22567	38.25444	10.87373	2.160928	24.94031	28.92671	2.446154
3	5.145352	18.07189	0.197253	1.792613	22.32001	20.16010	7.220604	20.25527	6.845967	10.83237	15.64818
4	2.327028	20.89021	3.015577	1.025711	25.13833	17.34177	10.03893	23.07359	4.027643	8.014045	18.46651
5	27.72588	4.508644	22.38328	24.37314	0.260521	42.74063	15.35993	2.325263	29.4265	33.4129	6.932345
6	15.20056	38.4178	20.54316	18.5533	42.66592	0.185809	27.56651	40.60118	13.49994	9.513541	35.99409
7	12.41719	10.80005	7.07458	9.064446	15.04818	27.43193	0.051229	12.98343	14.1178	18.1042	8.376351
8	26.05026	2.833021	20.70765	22.69752	1.415102	41.06501	13.6843	0.64964	27.75087	31.73728	5.256722
9	1.948574	25.16581	7.291179	5.301313	29.41393	13.06617	14.31453	27.34919	0.247959	3.738443	22.74211
10	5.332486	28.54972	10.67509	8.685225	32.79785	9.682263	17.69844	30.7331	3.631871	0.354532	26.12602
11	17.84909	5.368144	12.50649	14.49635	9.616267	32.86384	5.483138	7.551525	19.54971	23.53611	2.944443

Table.2 presents error values for another image in the matching process based on objects' signatures, which are applied on unclear image in Fig.2 (d). As in Table.1, all cell's values represent the DIF of Equation (5), and the least value indicates to exact match of objects in an image and the input one.

Clearly, as seen one mis-matching is found in second row column two; however this mis-matching is acceptable because the objects in second and third columns are so close to each other in shape.

Table.2. the matching process of objects' signatures

							
1	0.056373	9.911202	6.924778	17.52564	4.48623	4.075137	9.227283
2	7.716881	2.250695	0.735729	9.86513	12.14674	11.73564	1.566775
3	6.836034	3.131541	0.145117	10.74598	11.26589	10.8548	2.447622
4	16.82239	6.854813	14.87723	0.759622	21.25225	20.84115	7.538732
5	4.465828	14.4334	11.44698	22.04784	0.035971	0.447064	13.74948
6	3.9633	13.93088	10.94445	21.54531	0.466557	0.055463	13.24696
7	9.283804	0.683772	2.302652	8.298207	13.71366	13.30257	0.000148

As the same way in Table.1, and Table.2, all other objects have been selected based on their signatures and have achieved 96% in the matching process. On the other hand, by applying SURF on the same image with different input objects, some mis-matching are found if the input object has changed in his position or orientation, even so, this mis-matching has not happened with the proposed algorithm under the same constraints. Fig.4 illustrates SURF Work in an example for this mis-matching with the second object in second column of Table.1 by 100 strongest feature points.

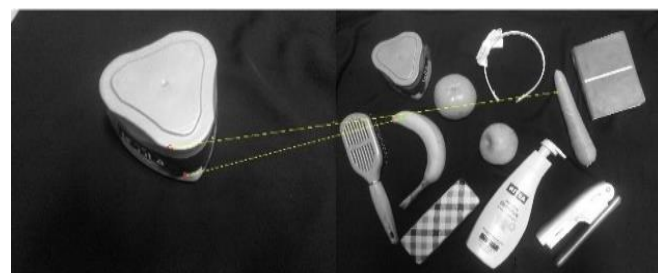
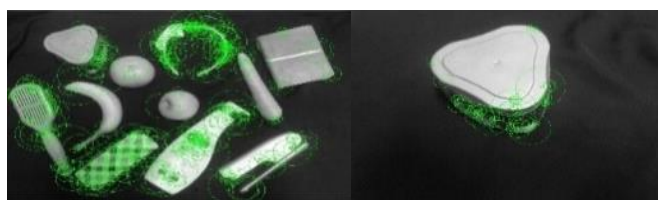


Fig.4. the SURF example with the mismatch an object

In this Figure, the input object in the left is mis-matched with its corresponding object in the original image. Additionally, this mismatch repeats many times with the test objects using SURF. From the previous results, although SURF method and proposed algorithm have presented to detecting and matching objects in an image, however, the presented algorithm is more effective and accurate in objects matching process than SURF and simply in use by some humble statistical equations without any constraints as in the other methods. Next section shows the conclusion of this work.

6 Conclusions

This paper have presented proposed algorithm for object detection and matching based on its own signature using morphological segmentation tools. The algorithm has divided into three parts; one is segmentation process, construction of object signatures, and the last part is the matching based on them to classify and define the object. Actually, this signature has a singularity, simply in use, saving it on a small memory, and working in a variety of light levels. Moreover, the proposed algorithm is matched the objects neither among object's blobs nor regions of interest; but among the constructed signatures. On the other hand, SURF method has been presented to comparison with the proposed method on the experimental results. Many difficulties are appeared in matching process, such as object in unusual intensity luminance, shape, orientation, position, different sizes, or unclear image of objects; but the proposed algorithm has overcome on them, while SURF has not do. The performance has been tested 120 from a wide variety of different objects, it has been achieved 100% in the case of constructing object signatures, also it has achieved 96% of exact matching whereas SURF has achieved 85% for all experimental objects.

7 Acknowledgment

This paper has been supported and funded by the deanship of scientific research in Salman bin Abdul Aziz University under a research projected 77/t/33h.

8 References

- [1]. Adel. A. Darwish., Hesham F. Ali., Hany A. M. El-Salamony. "3D Human Body Motion Detection and Tracking in Video". The 14th IASTED International Conference on Applied Simulation and Modeling, Spain, June 15-17, 2005.
- [2]. C. Papageorgiou and T. Poggio. "A trainable system for object detection". Intl. J Computer Vision, 38(1):15–33, 2000.
- [3]. C. Zahn and R. Roskies. "Fourier descriptors for plane closed curves". IEEE Trans. Computers, 21(3):269–281, March 1972.
- [4]. D. Gavrila and V. Philomin. "Real-time object detection for smart vehicles". In Proc. 7th Int. Conf. Computer Vision, pages 87–93, 1999.
- [5]. D. Huttenlocher, R. Lilien, and C. Olson. "View-based recognition using an eigenspace approximation to the Hausdorff measure". PAMI, 21(9):951–955, Sept. 1999.
- [6]. Elsalamony, H. A., "Automatic video stream indexing and retrieving based on face detection using wavelet transformation," *Signal Processing Systems (ICSPS), 2010 2nd International Conference on* , vol.1, no., pp.V1:153-157, 5-7 July 2010.
- [7]. Elsalamony, H. A., "Automatic object detection and matching based on proposed signature," *Audio, Language and Image Processing (ICALIP), 2012 International Conference on* , vol., no., pp.68,73, 16-18 July 2012.
- [8]. G. Bouchard and B. Triggs. "A hierarchical part-based model for visual object categorization". In CVPR, 2005.
- [9]. H. Bay, A. Ess, T. Tuytelaars, and L. VanGool, "Surf: Speeded up robust features," *International Journal of Computer Vision and Image Understanding* vol. 110, June 2008.
- [10]. H. Bay, T. Tuytelaars, and L. VanGool, "Surf: Speeded up robust features," presented at the Proceedings of European Conference on Computer Vision, Austria, 2006.
- [11]. Henry A. Rowley., Shumeet Baluja., and Takeo Kanade. "Human face detection in visual scenes". *Advances in Neural Info. Proc. Systems*, volume 8, 1995.
- [12]. Imtiaz Ali, Julien Mille, Laure Tougne, "Space-time spectral model for object detection in dynamic textured background". *Pattern Recognition Letters*, Volume 33, Issue 13, 1 October 2012, Pages 1710-1716, ISSN 0167-8655.
- [13]. Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox."Detection-based Object Labeling in 3D Scenes" *International Conference on Robotics and Automation*, 2012.
- [14]. Li He ; Hui Wang ; Hong Zhang . "Object detection by parts using appearance, structural and shape features". International Conference on Mechatronics and Automation (ICMA), page(s): 489 – 494, 2011.
- [15]. Lowe, David G, "Distinctive Image Features from Scale-Invariant Key points," *International Journal of Computer Vision* vol. 60, January 2004.
- [16]. Lowe, David G, "Object recognition from local scale-invariant features," presented at the Proceedings of the International Conference on Computer Vision, Greece, 1999.

- [17]. Luo Juan, Oubong Gwun, "A Comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing* vol. 3, June 2009.
- [18]. M. Fischler and R. Elschlager. "The representation and matching of pictorial structures". *IEEE Trans. Computers*, C-22(1):67–92, 1973.
- [19]. P. Viola, M. Jones, and D. Snow. "Detecting pedestrians using patterns of motion and appearance". In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
- [20]. R. C. Veltkamp and M. Hagedoorn. "State of the art in shape matching". Technical Report UU-CS-1999-27, Utrecht, 1999.
- [21]. Y. Amit, D. Geman, and K. Wilder. "Joint induction of shape features and tree classifiers". *IEEE Trans. PAMI*, 19(11):1300–1305, November 1997.
- [22]. Yan Ke, Rahul Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," presented at the 2004 Proc. of IEEE Conference on Computer Vision and Pattern Recognition, USA, 2004.

A Statistical Measuring System for Rainbow Trout

Marcelo Romero, José M. Miranda, Héctor A. Montes, Juan C. Acosta

Universidad Autónoma del Estado de México

[\[mromeroh, jmmirandac, hamontesv, jcacostag\]@uaemex.mx](mailto:[mromeroh, jmmirandac, hamontesv, jcacostag]@uaemex.mx)

Abstract

Traditionally, a manual method is used to classify the rainbow trout in small farms, which might cause stress and physical damage to the fish. Additionally, this manual classification is not always accurate, as farmers only visually check whether the trout is fry, fingerling or table-fish size. In this paper, we introduce a simple statistical model to measure rainbow trout in farms. For this research, we have designed and implemented a novel prototype that includes canalisation, illumination and vision components to take a 2D downward-view image of the trout. After that, this image is pre-processed to get the trout's contour, which is used to estimate the fish's length by adjusting the best regression curve to this contour. Finally, the trout's size is defined by the minimum Mahalanobis distance to training data. We have evaluated our experimental results as a binary classification problem and the best precision scores are 91.66% and 100% when classifying fingerling and table-fish trout, respectively.

1. Introduction

Generally, small farms use a manual measuring and counting process when cultivating rainbow trout [1-3], [11]. There are many reasons to perform such classification, but the most important are not only to feed the trout according to its size, but also to avoid cannibalism into the tanks [11]. Problems when doing a manual classification are, indistinctively, the stress and physical damage causes to the specimen when manipulated by the farmer. Moreover, it is believe that this classification approach is not accurate, where the trout is taken from the water using a net and visually the farmer decide whether or not the trout should be changed to another tank.

Mexico, as well as many other countries in the world, has large hydric areas which are ideal for aquaculture [5-6]. Taking advantage of both, its altitude and natural water resources, the State of Mexico (Mexico) has particular interest in increasing the trout's production as a sustainability and

economics strategy for local small farmers [7]. Hence, this is a good opportunity to integrate technology to optimise the trout's production in this region.

Hence, this motivated our research interest in the field, where we have accomplished some results, including a research project and bachelor in science final dissertations [1-3]. In this paper we report our experimental results' by practicing in a small farm located in the Valley of Toluca, Mexico [4], where we have observed a manual classification process as illustrated in Figure 1.

Some related work is observed in the literature. Ching-Lu et al. [14] proposed a technique to measure dead tuna fish using a colour pattern. In this work the fish length is estimated by proportional relationship between the fish body pixel length and the image reference scale. Nery et al. [15] measure four dead fish classes by constructing a central line along the fish body from horizontal and vertical views of the fish's body. Finally, a commercial counting and measuring system is observed in Vaky [16], however, there is no further information about its classification procedure.

The rest of this paper is as follows. Firstly, Section 2 describes our novel prototype designed for this research. Then, Section 3 introduces our statistical measuring approach. After that, Section 4 details our experimental framework. Next, Section 5 shows our performance evaluation. Finally, Section 6 concludes this paper and draws some venues for our future work.



Figure 1. Manual measuring-classification process generally done in small farms in Central Mexico. Note that this small farms use lined earth tanks.

2. Experimental prototype

In this Section we describe our experimental prototype, which has been designed as part of this research.

This novel prototype is essential to collect useful trout's 2D images; hence we have meticulously designed it. Figure 2, shows our experimental prototype, which has evolved from a traditional squared glass fish-cube (prototype version 1). We have observed relevant issues from our first prototype and that knowledge has been experimentally analysed to obtain our second model. Note that our two prototypes have been experimentally evaluated into a trout farm; therefore, we have gathered special knowledge about handling the rainbow trout.

Then, as observed in Figure 2, our experimental prototype consists of three main components: canalization, illumination and vision which are aim to collect 2D trout images using a standard personal computer.

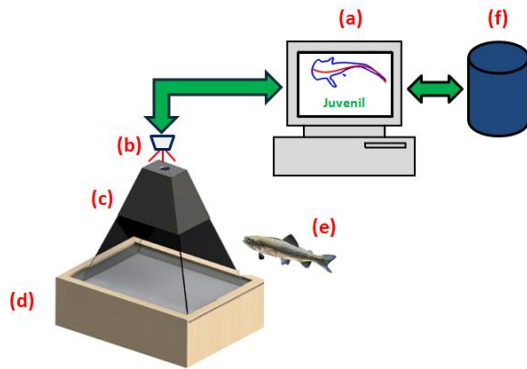


Figure 2. Our experimental scenario to measuring rainbow trout. (a) Statistical approach within a personal computer, (b) Vision system, (c) Canalisation system, (d) Illumination system, (e) Specimen to be measured, and (f) Database.

2.1 Canalisation system

We have design a novel canalisation system based on opaque-glass within our prototype. This canalization system poses two main properties. The first property is regarded to its trapezoidal shape, which has been considered according to the digital camera's vision field principle. As long as such trapezoidal shape avoids reflection to be captured when taking a digital image. In second term, we can mention that this is a two-canal tray, which prevents occlusion by taking only one fish per canal and it allows capturing two rainbow trout per shot.

2.2 Illumination system

To assist our vision system, we have integrated an illumination system, which distributes light in a

uniform way at the bottom of the canalisation system. To do this, a light source is located at according distance to distribute light uniformly over an acrylic diffuser. The light source's high was defined by using a bisection approach and measuring the light intensity projected into the diffuser with photo-resistors. We integrated this diffuse illumination to increase contrast into the image and highlighting the trout's body.

2.3 Vision system

In order to explore economical technology for our classification system, we have used a basic 2D WebCam® camera in this experimentation. This camera is able to capture RGB-images with a maximum resolution of 1920x1080 pixels. In this prototype, this 2D camera is located at the top of the canalization system to capture downward-view images of the trout. Its high is proportional to the canalization base length to avoid extra data to be captured.

3. Statistical measuring approach

In this Section we present our statistical approach to measure rainbow trout.

Considering the rainbow trout natural swimming movement against the water flow and observing the trout from a downward point of view, we hypothesised that a *third order curve* could approximate the trout's body within the water.

Different procedures can be followed to obtain a third order curve equation. However, we prefer a simple but effective solution that could be executed on-line after a trout's image is captured.

Then, given n sample points (x, y) which depict the trout body, we apply minimum squares to compute a polynomial third order equation [8]:

$$y^3 = a_0 + a_1x + a_2x^2 + a_3x^3 \quad (1)$$

Where, a_0, a_1, a_2, a_3 are constants that gain their values by solving the [4x4] equation system (2):

$$\sum_{i=1}^n y_i = na_0 + a_1 \sum_{i=1}^n x_i + a_2 \sum_{i=1}^n x_i^2 + a_3 \sum_{i=1}^n x_i^3$$

$$\sum_{i=1}^n x_i y_i = a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 + a_2 \sum_{i=1}^n x_i^3 + a_3 \sum_{i=1}^n x_i^4$$

$$\sum_{i=1}^n x_i^2 y_i = a_0 \sum_{i=1}^n x_i^2 + a_1 \sum_{i=1}^n x_i^3 + a_2 \sum_{i=1}^n x_i^4 + a_3 \sum_{i=1}^n x_i^5$$

$$\sum_{i=1}^n x_i^3 y_i = a_0 \sum_{i=1}^n x_i^3 + a_1 \sum_{i=1}^n x_i^4 + a_2 \sum_{i=1}^n x_i^5 + a_3 \sum_{i=1}^n x_i^6$$

The equation system (2) can be easily solved using the matrix notation, $AX = B$, or more specifically:

$$X = BA^{-1}.$$

After this computation, we obtained the best regression curve that adjusts the trout's body captured into a 2D image.

Then, we observe that this regression curve is related to the trout's length, which could be estimated by computing the Euclidean distance among the points within the regression curve.

Finally, given a probe-length (l_i) a classification can be done by comparing against training lengths. For this research, such comparison is performed by computing the Mahalanobis distance [10] from training fry, fingerling and table-trout lengths:

$$d_i = \frac{l_i - \bar{x}}{s_{\bar{x}}} \tag{3}$$

Hence, a probe-trout t_i is classified through its length l_i by comparing its Mahalanobis distance d_i against a predefined threshold, which in fact is the number of standard deviations that is expected to be l_i to the training main (\bar{x}).

4. Experimental framework

This section presents the experimental framework to illustrate how rainbow trout is measured using our statistical approach.

As show in Figure 3, after an RGB image is taken by our prototype, we are following a five stages image processing to get the trout's contour. As explained in Section 3, we are measuring the trout's length using this contour.

To classify the trout's image within an image, we are performing four main steps. First, an RGB image of the trout is taken using our prototype (Section 2). Second, this RGB image is processed to obtain the trout's contour. Third, the trout's length is estimated by applying our statistical method (Section 3) to the trout's contour. Finally, using that estimated length, the trout is classified using a binary classification approach.

In Subsection 4.1, we provide more detail about our image processing step.

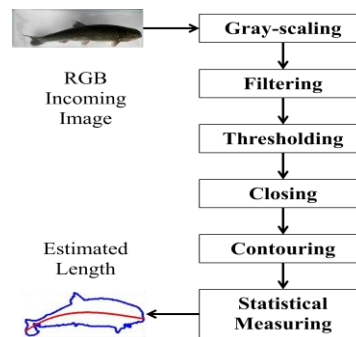


Figure 3. Processing an incoming image to estimate the trout's length using our statistical approach.

4.1 Testing procedure

As illustrated in Figures 2, we have implemented a novel functional prototype which allows us to gather RGB images. After that, as shown in Figure 3, RGB images are processed (using standard algorithms in the literature [9], [12]) until we obtain the trout's contour. Next, we apply our statistical approach to that contour, so we can estimate the trout's length. Finally, a binary classification approach is taken to classify the trout within the income image. We detail our experimental procedure:

1. As illustrated in Table I, for this experiment, we have collected a trout-image database using our prototype in a farm (illustrated in Figure 4). This database was created using 30 fingerling and 8 table-fish specimens, capturing 8 images per specimen. We regret that on this visit we were unable to experiment with fry trout, because of the season.

TABLE I. EXPERIMENTAL DATA IMAGES.

Trout Size	# Specimens	# Images per Specimen	Total images
Fingerling	30	8	240
Table-fish	8	8	64
Grand total	38		304

2. From our database, separate training and testing sets are defined (see Table II). Thus, 38 images for training and 266 images for testing are used. Specifically, we have two training data, one for fingerling size (30 images) and another for table-fish size (8 images). In both cases, we selected the first captured image to be part of the training set. Hence, we have 210 fingerling images and 56 table-fish images for testing.

TABLE II. TRAINING AND TESTING SETS.

Trout Size	Training	Testing
Fingerling	30	210
Table-fish	8	56
Total	38	304

3. From these 30 and 8 training images, training data is gathered, which in fact consist of training lengths, the arithmetic mean and the standard deviation for each size.
4. For each testing trout image, estimated length are gathered as illustrated in Figures 3 and 5. To do this, we gather an RGB image using our prototype. Then, we execute a five stages image processing: gray-scaling, filtering, thresholding, closing, and contouring. Next, we estimate the trout's length by apply our statistical approach to the contour obtained above. Finally, using this estimated length we classify the trout as fingerling or table-fish.
5. To speed up our image processing step, our vision system gathers 640x360 pixels RGB-images.
6. Captured RGB values are converted into a grayscale by forming weighted sums of the R, G, and B components:

$$0.2989 * R + 0.5870 * G + 0.1140 * B \quad (4)$$

7. Noise reduction is performed in every grayscale image by using a [3x3] Gaussian lowpass filter and $\sigma = 0.5$

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}} \quad (5)$$

8. A binary image is obtained by using a 0.245 threshold, which was calculated experimentally from training rainbow trout images.
9. The trout's body is emphasized by using a closing operation, first erosion and then dilation with a [5x8] mask. This operation is the key to eliminate small clusters of pixels around the trout's body cluster.
10. The trout's contour is obtained by removing interior pixels. In this case, a pixel is set to 0 if all its 4-connected neighbours are 1, thus leaving only the boundary pixels on:

If

	1	
1	x	1
	1	

 Then

	1	
1	0	1
	1	

 (6)

11. Using the trout's contour, we apply our statistical measuring approach detailed in Section 3.

12. By definition, the trout's size is estimated by computing the Mahalanobis distance from this estimated length to training data (Eq. 3).
13. For the classification in this experiment, imagine that the complete testing-trout set (266 in total) is passed through out a grid one by one in two steps. Firstly, the grip is sized to filter only fingerling. Then, every trout able to pass this grid is a fingerling. Secondly, for the rest of the testing set, the grid is now sized to filter table-fish trout.

Remember that we are computing the Mahalanobis distance and this allow us to easily implement the approach above by using firstly fingerling training data and secondly table-fish training data. Referring as training data the arithmetic mean and the standard deviation from each size.

Another advantage in using Mahalanobis distance, is that we can make our classification process as rigid as we decide, by defining a threshold in number of standard deviations.

Then, in this paper we are reporting classification figures from one to six standard deviations.

14. We are considering this experiment as a binary classification problem, as illustrated in Table III. By doing this, we are collecting true positive (TP), false positive (FP), false negative (FN) and true negative (TN) frequencies [13].

TABLE III. BINARY CLASSIFICATION.

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

15. Using values in Table III, performance figures are generated by computing accuracy, repeatability, and specificity when classifying as fingerling and table-fish trout. Furthermore, we are presenting a recall-precision curve for our classification procedure. In every case, we are evaluating using as threshold from one to six standard deviation.

Accuracy, a degree of veracity, is a measurement of how well the binary classification test correctly identifies a rainbow trout's size.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Repeatability, a degree of reproducibility, is an indicator about how robustly a rainbow trout size can be identified.

$$\text{Repeatability} = \frac{TP}{TP + FP} \quad (8)$$

Specificity, a degree of speciality, rates how negative rainbow trout's size is correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (9)$$

Recall measures the fraction of positive examples that are correctly labelled:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

Precision measures that fraction of examples classified as positive that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

5. Performance evaluation

We now present performance figures when using our statistical model to measure rainbow trout in farms.

As observed in Figure 5, our statistical approach's performance to measure a rainbow trout depends on our image processing stage. However, according to our experimental results, we believe that we have addressed main issues about capturing and processing an RGB image within our system.

As we have mentioned before, we consider this as a binary classification problem. To proceed with this, we are following step 13 in our experimental procedure (Section 4.1). Thus, the complete testing lengths (266 images) are compared against fingerling training data, using Mahalanobis distance. Then, if a testing length falls into a predefined threshold (one to six standard deviations) the respective testing trout is marked as fingerling. Next, all remaining testing lengths are compared against table-fish training data using Mahalanobis distance as well. Again, if the testing length falls into a predefined threshold (one to six standard deviations) we label the respective testing trout as table-fish trout.

Then, as prescribed in Table III we count TP, FP, TN and FN frequencies, which are summarised in Tables IV and V. Hence, by using these values we are able to compute accuracy, repeatability, and specificity metrics, which is presented in Table VI and illustrated in Figures 6 to 8.

Finally, we are computing recall and precision metrics. Tables VII and VIII summarises those results and Figure 9 plots the respective recall-precision curve.

Observing our experimental results when classifying our testing set as fingerling, we score the best precision, 91.66% at one standard deviation. Whereas, our best recall score, 94.28%, is obtained at six standard deviations.

Furthermore, when classifying those testing lengths that do not fell into a fingerling class, we scored a 100% precision and recall, in one and three standard deviations, respectively.

These experimental results are not only motivating, but also a valuable evidence that indicates effectiveness in our classification system.



Figure 4. Functional prototype used within our measuring system. This prototype includes an illumination source, a pyramidal canalisation compartment and a 2D camera.

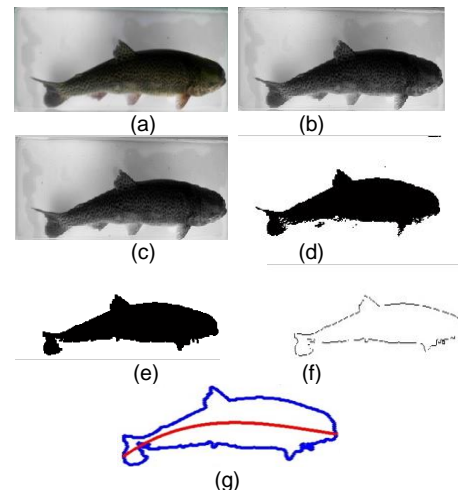


Figure 5. Image processing performed to measure a rainbow trout using our statistical approach. (a) RGB incoming image sensed by the vision system; (b) Gray-scaling; (c) Filtering; (d) Thresholding; (e) Closing; (f) Contouring; (g) trout's length estimated by a third order regression curve (plotted in red).

TABLE IV. FREQUENCY WHEN CLASSIFYING A PROBE SET AS FINGERLING.

	# Standard deviations					
	+/-1	+/-2	+/-3	+/-4	+/-5	+/-6
TP	110	166	178	190	195	198
FP	10	31	52	56	56	56
TN	46	25	4	0	0	0
FN	100	44	32	20	15	12

TABLE V. FREQUENCY WHEN CLASSIFYING A PROBE SET AS TABLE-FISH.

	# Standard deviations					
	+/-1	+/-2	+/-3	+/-4	+/-5	+/-6
TP	9	24	4	0	0	0
FP	0	0	0	0	0	0
TN	100	44	32	20	15	12
FN	37	1	0	0	0	0

TABLE VI. CLASSIFICATION SUMMARY.

Classification Stage	Testing images	Accuracy	Repeatability	Specificity
Evaluating as fingerling	266	74%	92%	82%
Evaluating as table-fish	110	100%	100%	100%

TABLE VII. RECALL-PRECISION WHEN CLASSIFYING A PROBE SET AS FINGERLING.

	# Standard deviations					
	+/-1	+/-2	+/-3	+/-4	+/-5	+/-6
Recall	52.38%	79.04%	84.76%	90.47%	92.86%	94.28%
Precision	91.66%	84.26%	77.39%	77.23%	77.69%	77.95%

TABLE VIII. RECALL-PRECISION WHEN CLASSIFYING A PROBE SET AS TABLE-FISH.

	# Standard deviations					
	+/-1	+/-2	+/-3	+/-4	+/-5	+/-6
Recall	19.56%	96.00%	100%	0%	0%	0%
Precision	100%	100%	100%	0%	0%	0%

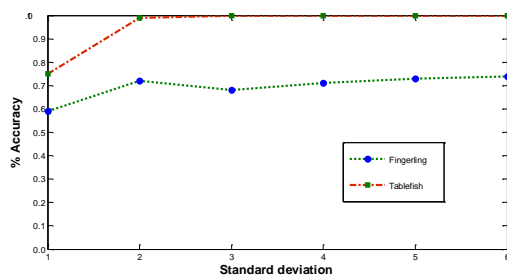


Figure 6. Accuracy performance when classifying young and adult rainbow trout using our statistical model.

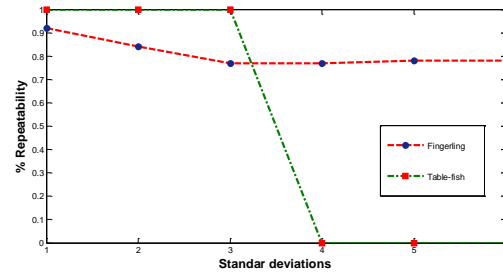


Figure 7. Repeatability performance when classifying young and adult rainbow trout using our statistical model.

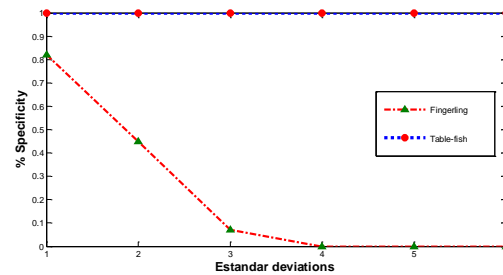


Figure 8. Specificity performance when classifying young and adult rainbow trout using our statistical model.

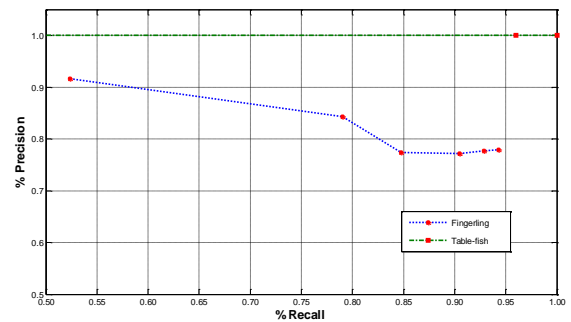


Figure 9. Recall-precision curves when classifying fingerling and table-fish trout.

6. Conclusions

In this paper we have introduced our statistical system to measure rainbow trout in farm using computer vision. This novel technique is a simple but effective statistical method which has been evaluated in a small farm in Central Mexico.

For this research, we have designed and implemented a functional prototype to collect 2D trout images. This prototype includes canalization, illumination and vision components which have been meticulously assembled. Also, we believe that this prototype could be easily integrated into a mechanical system to interconnect lined earth tanks in farms.

Our preliminary results encourage our research as they have shown that our classification system is

effective, where a 91.66% and 100% precision are observed when classifying fingerling and table-fish trout, respectively.

It is important to observe that, although our statistical approach has been inspired to measure rainbow trout, this approach can be applied to other fishes grown in farms.

As part of our future work we are integrating a water flow into our prototype as well as a stereo vision system, then, we are investigating two main issues: the light reflection into the water and the presence of turbulence. Our final aim is to implement an economical classification system which would be installed in small farms in Central Mexico.

References

- [1] Romero, Vilchis and Portillo (2012). Intelligent system to count, measure and classify fishes using computer vision. Research project StyEA 32742012M. Autonomous University of the State of Mexico.
- [2] Serena Mejía-Pichardo (2013). Cuento y clasificación de la trucha arcofiris utilizando visión artificial: revisión literaria y análisis. BSc. Final Dissertation. Engineering Department. Autonomous University of the State of Mexico.
- [3] José Manuel Miranda-Contreras (2014). Prototipo de un sistema clasificador de la trucha arcofiris utilizando un modelo estadístico de su longitud obtenido de imágenes 2D. BSc thesis, Engineering Department, Autonomous University of the State of Mexico.
- [4] Rincon del sol (2014). Small Trout Farm. La Marqueza, State of Mexico, Mexico.
- [5] Comisión Nacional del Agua (2006). Concesión de Aprovechamiento de Aguas Superficiales. Water National Council.
- [6] Diario Oficial de la Federación (2004). Ley de aguas nacionales. DCVII(14): 27 – 95. Federal Official News.
- [7] Gallego A. I., R. Carrillo, D. García, L. Sasso, J. Guerrero, R. Carrillo, D. García, C. Díaz, C. Fall, C. Burrola, L. White, J. Manjarrez, C. Zepeda, X. Aguilar, G. Legorreta y A. Sánchez. (2007). Programa maestro, sistema producto trucha del estado de México. Government Plan for Rainbow Trout Production. Autonomous University of the State of Mexico, México.
- [8] Murray Spiegel (2012). Probabilidad y Estadística. Mc Graw-Hill.
- [9] Rafael Gonzalez and Richard Woods (2008). Digital Image Processing. Prentice Hall.
- [10] Richard Duda, Peter Hart, David Stork (2001). Pattern Classification. Wiley Interscience.
- [11] András Woynarovich, György Hoitsy and Thomas Moth-Poulsen (2011). Small-scale rainbow trout farming. FAO Fisheries and aquaculture technical paper (561), Food and Agriculture Organization of the United Nations.
- [12] Rafael Gonzalez, Richard Woods and Steven Eddins (2004). Digital image processing using Matlab. Prentice-Hall.
- [13] Jese Davis and Mark Goadrich (2006). The relationship between precision-recall and ROC curves. In proceedings of the 23rd International Conference on Machine Learning.
- [14] Ching-Lu Hsieh, Hsiang-Yun Chang, Fei-Hung Chen, Jhao-Huei Liou, Shui-Kai Chang, Ta-Te Lin (2011). A simple and effective digital imaging approach for tuna fish length measurement compatible with fishing operations. Computers and Electronics in Agriculture, Volumen 75.
- [15] Ibrahim, M.Y., Wang, J (2009). Mechatronics Applications to Fish Sorting Part 1: Fish size identification. Industrial Electronics (ISIE 2009). IEEE International Symposium.
- [16] Vaki (2013). Bioscanner Fish Counter. <http://www.vaki.is/Products/BioscannerFishCounter/>

Highlight Image Filter Significantly Improves Optical Character Recognition on Text Images

Iulia Ştirb

Computer and Software Engineering Department, "Politehnica" University of Timișoara, România

Abstract - Image filtering is the change of the appearance of an image by altering the shades and colors of the pixels. Increasing the contrast as well as adding a variety of texture, tones and special effects to images are some of the results of applying filters. In order to obtain a high success rate of OCR (Optical Character Recognition) on images which contain text, the main target of filters is, however, reducing the noise of the image. Within Silicon and Software System Limited (S3Group) Company from Dublin, Ireland, the OCR is part of the process of testing the menu options displayed on the output video of the set-top box under test. The important role played by image filters in improving a subsequent OCR process on text images was the reason I created two new non-linear efficient filters that will be presented in this paper. The first image filter that is named Smart Contrast, increases the contrast of the image in a way depending on the value of each component (Red, Green and Blue) of each pixel in the image. The second image filter, called Highlight, produces an outstanding increase of OCR success rate on the filtered images. As Highlight filter is carried out, the implementation differs from all other known filters, while the visual effect on the filtered images can be described as a combination between increased contrast as said before with other two visual effect: sharpening and highlighting the edges (of the characters). Precisely, combining these specific visual effects on the resulting image makes Highlight filter so powerful in improving OCR on text images.

Keywords: image, filter, highlight, sharpen, contrast, OCR

1 Introduction

Many often, the choice of the image filter is done non-automatically, by humans, as a result of observing the characteristics of the image (color, shape or thickness of text in the image). Researches reveal that specific filters are used for certain images. For instance if the original image is blurred and the expected result is an image with a higher clarity than Sharpen filter can be used and on the other hand, if a less level of details is desired in the resulting image, than Blur filter would be the right choice.

As in [2], the process of selecting the scale of a filter in order to perform edge detection over the image can be automated. Overall, as in [1], researches have been carried out regarding the automated

selection of the filters' parameters. In other words, once the proper filter to apply to the image has been choose by humans, the filter parameter is selected by a computer depending on the desired output image.

However, an automated analysis of the image properties in order to select the proper filters to be applied to it would be a complicated, expensive and time consuming process since the analysis depends on many factors (e.g. noise, clarity or contrast of the input image).

The image filter which I named it "Highlight" is designed to be a universal filter for improving OCR rate of success on a large variety of text images and because of its large applicability it avoids the automated selection of the proper filter to be applied to a specific image.

The paper contains one major section in which Highlight new image filter and some already existing filters are described. Snippets of code from Highlight filter implementation are also shown in this section. Conclusions section presents the major benefit Highlight image filter brings in improving the success rate of OCR in comparison to other filters and describes the visual effect of Highlight filter on images.

1.1 Properties of Highlight image filter

Highlight image filter detects the edges of the features in the image i.e. edges of the characters in a text image, highlights and sharpens the text and increases the contrast of the edges of the characters in a way similar to Smart Contrast image filter.

What Highlight filter brings new regarding the way contrast is done is that it performs a selection between two types of transformation and choses the proper one to be applied. Instead of simply applying the same transformation to all components of each pixel like Contrast filter does, the selection is done for each component (e.g. Red) of each pixel. This improvement made by Highlight image filter regarding the way contrast is increased produces even more contrast between the text in an image and the background in cases when this is needed such as when the colour of the text is close to the colour of the background making the text more visible than it would be by simply applying Contrast filter.

2 Description of Smart Contrast image filter

Smart Contrast filter compares the value of each component (e.g. Red) of each pixel with 127 (255 / 2), note that the range in which the components Red, Green

and Blue vary is 0 - 255) and if the value is less if less than 127 perform a certain transformation to that specific component, if greater perform a different transformation. As a case study, if it is considered an image which contains some text and the color of the text would be $(R_t, G_t, B_t) = (126, 126, 125)$ and the color of the background would be $(R_b, G_b, B_b) = (130, 137, 136)$, then the colors would be pretty similar, so the text would be hard to recognize even for the human eye. In this case, Smart Contrast filter decreases more the color of the text and increases more the color of the background than Contrast filter would do, making the text more visible and more easily to be detected. Thus, Smart Contrast keeps the good work Contrast filter does and, in addition, produces good results for edge cases.

2.1 Important step in creating Smart Contrast filter: knowing how Contrast filter works

Contrast filter is based on the transformation in (1) where *contrast* is the contrast scale (the degree to which the contrast is increased) and *red*, *green* and *blue* are the values of the components of a pixel.

$$((red / 255.0 - 0.5) * contrast + 0.5) * 255.0 \quad (1a)$$

$$((green / 255.0 - 0.5) * contrast + 0.5) * 255.0 \quad (1b)$$

$$((blue / 255.0 - 0.5) * contrast + 0.5) * 255.0 \quad (1c)$$

The graphic representation for transformation in (1) is represented in Fig. 1 (blue plot). In the same figure, the identity function it is also drawn (green plot) to spot how pixel components values increase or decrease according to the transformation. If the value of the pixel component is less than zero it is set to zero and if it is greater than 255 it is set to 255.

So far, the same transformation is being applied to all components of each pixel. Thus, this is how Contrast filter performs, however the property is also specific to many other filters (e.g. Invert, Color).

2.2 New image filter: Smart Contrast

Smart Contrast performs two similar transformations depending on the value of the pixel component. For less than 127 (255 / 2) values the formula is illustrated by:

$$((value/255.0-0.6)*contrast+0.6)*255.0 \quad (2)$$

Equation (3) shows the formula for values of the pixel components greater than 127.

$$((value/255.0-0.4)*contrast+0.4)*255.0 \quad (3)$$

Value 127 is the threshold for Smart Contrast filter and the value of the threshold is chosen to be the median value from the range 0 to 255 (i.e. 0 is lowest

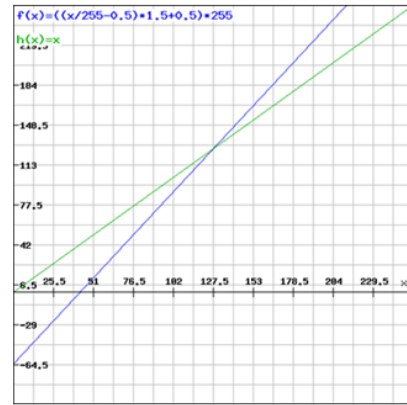


Figure 1. Contrast filter transformation (bleu plot)

value and 255 the highest value for color intensity).

As a remark, the same pixel could be the result of applying two types of transformations to its components (e.g. Red, Green or Blue). For instance, we could focus on an arbitrary pixel that has the coordinates (x,y) relative to the upper-left corner of the image. By assuming that (2) is applied to the Red component of that pixel and (3) is applied to the Blue component of the same pixel, we are facing a possible situation that could arise in the algorithm. Despite this fact exemplified before, no more than one transformation will be applied to a single component of a pixel (e.g. Red component could not possibly be the result of applying (2) and (3), it will have to be either (2) or (3), but not both).

Furthermore, two different pixels could be the result of applying different transformations to the same component of the two pixels (e.g. the filtered Red component of the first pixel that has the coordinates (x_1, y_1) could be the result of applying (2) and the filtered Red component of the second pixel that has the coordinates (x_2, y_2) could be the result of applying (3)). The graphic representation of the two transformations is shown in Fig. 2A, together with the identity function that helps in spotting the way pixel components are increased or decreased. Fig. 2B highlights the difference between Contrast and Smart Contrast algorithms. If the value of the pixel component is less than threshold 127 the blue plot describes the transformation that is applied to that certain component, else the transformation shown in the red plot is the one applied to the component.

Best OCR rate of success for the filtered images using Smart Contrast filter is produced when contrast scale is set to 1.5.

2.3 Visual result of applying Smart Contrast on images

Smart Contrast filter produces the results shown in Fig. 3. The results produced by Contrast filter are also shown in Fig. 3 in order to spot the improvements made by Smart Contrast. The effect of applying Smart Contrast filter would be that, in most of the cases, contrast is increased in the areas of the image where characters appear. Exceptions occur when the color of

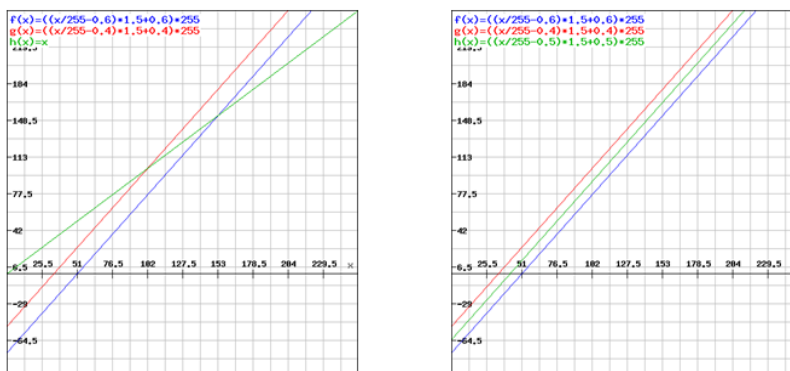


Figure 2. A. Smart Contrast transformations (blue and red plot) and the identity function (green plot)
 B. Smart Contrast (blue and red plots) and Contrast transformations (green plot)

the characters is close to the color of the background and both are close to either the lowest color intensity either the highest color intensity. This drawback is solved in Highlight filter.

3 Highlight image filter description

This filter decreases more the values of pixel components (i.e. Red, Green and Blue) that are less than 127 (using (2)) and increases more the values greater than 127 (using (3)) than Contrast filter would do. OCR benefits from Highlight filter's improved way of contrasting the image (which is similar to the way Smart Contrast filter performs) and from the other two properties that are sharpening and highlighting the edges of the features in the image (e.g. the characters).

Highlight filter detects the areas of rapid intensity change (i.e. edges) like Laplacian of Gaussian filter would do. Once the edges are detected, they are being sharpened, which would produce a visual effect that is similar to what Sharpen filter would do to an image. Still, the implementation of Highlight image filter has no similarities with Sharpen and Laplacian of Gaussian filters' implementation.

In addition, Highlight filter creates shadows behind characters (the color of the shadows contrasts with the color of the characters).

All this combined properties of contribute to a better success rate of OCR on text images.

3.1 Important step in creating Highlight image filter: understanding the effect of Sharpen and Laplacian of Gaussian filters on images

Smart Contrast produces an effect similar to what Contrast does, that is increasing the contrast of



Figure 3. The visual effect of Smart Contrast image filter

the image. In addition, Smart Contrast decreases more the values of pixel components that are less than 127 and increases more the values greater than 127 than Contrast would do. Highlight filter contrasts the image in a way similar to Smart Contrast.

Sharpen filter accentuates edges, but it does as well with the noise, as in [3], which is undesired and could make the OCR produce worse results than with the unfiltered image. Highlight filter takes the concept of spotting the edges from Sharpen filter, but does not accentuate the noise.

Laplacian of Gaussian combines the effects of Laplacian filter and Gaussian filter (which blurs the images to reduce the sensitivity to noise). While Laplacian detects the regions of rapid intensity change therefore being used in edge detection, Laplacian of Gaussian sharpens edges between two regions of uniform color but different intensities, as in [4].

3.2 New image filter: Highlight

Highlight filter gather together visual effect similar to the ones produced by Contrast, Sharpen and Laplacian of Gaussian filters. The implementation is carried out in an original manner using no template convolution (masks) like the last two filters do.

Highlight filter performs a different contrast increase for each component of each pixel in the image. A component - let's take as an example the Red component - of the current pixel is increased or decreased depending on the value of the Red component of the current pixel and the two other filled with Red color pixels as in Fig. 4 which shows a small 3 x 3 area within an image. The pixels which are not filled with Red color in the same figure do not contribute to the new value of the Red component of the current pixel.

Pixels which contribute to the new value of the current pixel in Fig. 4 are placed in a diagonal manner. Because of this, both vertical and horizontal edges of the characters are detected.

For each component of each pixel a contrast scale is computed separately. The way contrast scale of each component of the current pixel (x,y) is computed, is emphasized in (4), where r, g and b

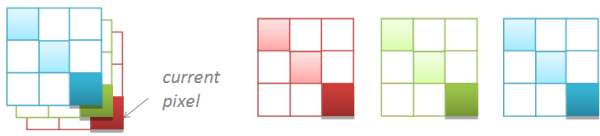


Figure 4. Pixels (filled with color) which count in computing the value of the contrast scale of each individual component (i.e. Red, Green and Blue) of the current pixel

indicate the Red, Green and Blue components and x_1y_1 , x_2y_2 indicate the pixels with coordinates $(x - 1, y - 1)$ and $(x - 2, y - 2)$. For instance, $r_{x_2y_2}$ is the value of the Red component of the pixel with coordinates $(x - 2, y - 2)$.

$$(100 + |r_{x_1y_1} - r_{xy}| + |r_{x_2y_2} - r_{x_1y_1}|) / 100 \quad (4a)$$

$$(100 + |g_{x_1y_1} - g_{xy}| + |g_{x_2y_2} - g_{x_1y_1}|) / 100 \quad (4b)$$

$$(100 + |b_{x_1y_1} - b_{xy}| + |b_{x_2y_2} - b_{x_1y_1}|) / 100 \quad (4c)$$

Equations (4a-c) produce values in the range from 1.0 to 6.1 and are not applied to the left and top edges of the image. Once the contrast for each component of each individual pixel in the image has been recorded, the algorithm is ready to be applied.

The starting point of Highlight filter algorithm is based on the fact that human eye is sensitive to a difference of at least 30 between the values of at least one of the same component of two adjacent pixel when it comes to perceive and recognize characters. To be more specific, if we would have to write some characters on a background which is uniformly colored with the intensity $(R_b, G_b, B_b) = (0, 0, 0)$, the color of the text would have to be $(R_t, G_t, B_t) = (0 + 40, 0 + 31, 0)$ or $(R_t, G_t, B_t) = (0, 0 + 31, 0)$ or any other combination that would meet the above request, in order for the human eye to recognize what is written. Fig. 5 proves what is being said before.

A reasonable assumption that is made from the start is that, characters that would not be perceived by the human eye, are not expected to be recognized by a machine using OCR, but every character perceived by the human eye must be also recognized by a machine (assuming there is no noise), or at least expect to be recognized. Thus, for the text that could not be perceived by the human eye, the performances of OCR are not improved by filtering first the text image using Highlight filter.

The core of the algorithm is described in the following. Since an example makes the general case more explicit, let's bring into discussion the Blue component of the pixel having the coordinates (x, y) relative to the upper-left corner of the image.

If the absolute difference between the value of the Blue component of the pixel with coordinates $(x - 1, y - 1)$ and the value of the Blue component of the pixel having the coordinates (x, y) is greater than 15 and the absolute difference between the value of the Blue component of the pixel with coordinates $(x - 2, y - 2)$ and the value of the Blue component of the pixel with coordinates $(x - 1, y - 1)$ is also greater

than 15 than a certain transformation will be applied to the Blue component of the pixel having the coordinates $(x - 2, y - 2)$ (for x and y greater than 2).

The requests described before and shown in (5) will be as well tested separately for the other two components i.e. Red and Green of each pixel in the image, except for the pixels in the right and bottom edges of the image (those will not be filtered). This remark is valid for (5), but as well for all the following formulas.

$$|r_{x_1y_1} - r_{xy}| > 15 \ \&\& \ |r_{x_2y_2} - r_{x_1y_1}| > 15 \quad (5a)$$

$$|g_{x_1y_1} - g_{xy}| > 15 \ \&\& \ |g_{x_2y_2} - g_{x_1y_1}| > 15 \quad (5b)$$

$$|b_{x_1y_1} - b_{xy}| > 15 \ \&\& \ |b_{x_2y_2} - b_{x_1y_1}| > 15 \quad (5c)$$

Why 15 (30/2)? Let's assume that the text in an image has some noise around it and the transition between the color of the text and the color of the background is done through an intermediary pixel which could be called "noise pixel" which has a different color. This situation appears frequently in the real scenarios. Then, the minimum difference of 30 between the values of the color component of the background and the color component of the text could be spread among three adjacent pixels (place as the colored ones in Fig. 4) with three different colors instead of just two adjacent pixels i.e. "text pixel" and "background pixel" as in Fig. 5. For instance, if the intensity of the color of the background would be $(R_b, G_b, B_b) = (0, 0, 0)$, the intermediate color (noise color) would have to be at least $(R_n, G_n, B_n) = (0, 0, 16)$ and the text color would have to be at least $(R_t, G_t, B_t) = (0, 0, 32)$, for the condition in (5c) to be fulfilled.

Because of the diagonal manner in which the pixels which contribute to the new value of the current pixels are placed, meaning that the direction of the gradient is a diagonal direction, all the noise around curve edges that follow this direction is eliminated. Noise around vertical and horizontal edges, which form an angle of -45° and 45° with the diagonal direction, is as well almost eliminated. Anyway, the noise affects more the curve edges than the horizontal and vertical edges.

Fig. 6 shows the visual representation of all possible cases that fulfill the condition in (5c), which refers to the Blue component. The visual representation for the Red and Green component can be obtained in the same way.

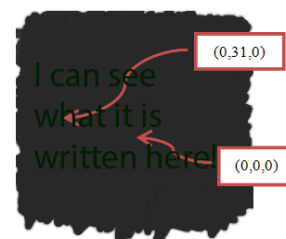


Figure 5. The condition to be met for characters to be recognized by human eye



Figure 6. Possible cases for the condition in formula (5c) to be fulfilled

Before getting to the point where a specific transformation is applied, another condition, in addition to the one described in (5c), must be first met. The new condition is described in (6c).

If the conditions in (6c) are fulfilled, there is a diagonal “blue” gradient. If (5c) and (6c) are met then the transformation described in the first paragraph of Section II is applied: to the Blue component of the current pixel so, therefore, the contrast for the Blue component is increased depending on the result of the comparison with 127 (255/2).

$$|r_{x_2y_2} - r_{xy}| > 30 \quad (6a)$$

$$|g_{x_2y_2} - g_{xy}| > 30 \quad (6b)$$

$$|b_{x_2y_2} - b_{xy}| > 30 \quad (6c)$$

Fig. 7 spots the two of the six possible cases shown in Fig. 6, showing the visual representation of the situations when condition in (6c) is met.

If the condition in (6c) is not fulfilled than the intensity of the Blue component will be increased or decreased depending on whether the condition in (7c) described below is met or not. To be more specific, if the condition in (7c) is fulfilled, the intensity of the Blue component is decreased using a transformation in (2) and in the opposite case the intensity is increased using transformation in (3). Both transformations are applied, this time, regardless of the value of the Blue component.

$$|r_{x_2y_2} - r_{x_1y_1}| > 0 \quad (7a)$$

$$|g_{x_2y_2} - g_{x_1y_1}| > 0 \quad (7b)$$

$$|b_{x_2y_2} - b_{x_1y_1}| > 0 \quad (7c)$$

Transformations in (2) and (3) are applied to the value of each component i.e. value variable and the result depends on the contrast scale i.e. contrast variable.

Fig. 8 spots the two of the 4 left possible cases shown in Fig. 6 (the first two of them were already discussed and matched with the condition in (7c)). More exactly, Fig. 8 shows the visual representation of the cases which fulfill the condition in (7c).

In case the condition in (7c) is not fulfilled, the two left cases that were not discussed before, out of six spotted in Fig. 6, are shown in Fig. 9.

It was never said before what happens if the condition (5c) is not met and this will be the appropriate time to be speaking about this. Well, if the condition is not fulfilled another condition is being tested and shown in (8c).

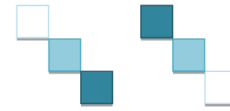


Figure 7. Possible cases for the condition in (6c) to be fulfilled



Figure 8. Possible cases for the condition in (7c) to be fulfilled

$$|r_{x_2y_2} - r_{x_1y_1}| > 15 \quad (8a)$$

$$|g_{x_2y_2} - g_{x_1y_1}| > 15 \quad (8b)$$

$$|b_{x_2y_2} - b_{x_1y_1}| > 15 \quad (8c)$$

If neither (8c) is fulfilled no transformation will be applied to the Blue Component of the pixel with coordinates (x-2,y-2). If (8c) is met, the condition in (7c) will be tested again. The visual representation of the cases that meet the condition in (8c) is shown in Fig. 10.

3.3 Visual results of applying Highlight filter on images

Highlight image filter produces the results in Fig. 11. It can be seen how this filter detects the edges of the characters and sharpens them (the best example would be “ALL CHANNELS” image) and how it creates contrasting shadows behind the characters (the black shadows can be best seen on white colored “Golf: Women’s British Open” text in the image).

Because of the shadows behind the sharpened edges of the characters and because of the increased contrast in the edges, the characters appear to be highlighted in the filtered image, which is the main visual effect of Highlight image filter.

3.4 Highlight image filter program code and visual representation

Part of the C# code that corresponds to Highlight image filter is being listed in Code 1. *buffer* array stores the image representation, more exactly the Blue, Green, Red and Alpha components in this order for the first pixel, then the components for the second pixel in the same order and so on for the rest of the pixels in the image.

Code 1. A part of the Highlight image filter algorithm

```
public void EdgeIntensityChange(byte[] buffer, double[]
    contrastBuffer, int Stride, int k)
{
    int diff01 = buffer[k - Stride * 2 - 8] - buffer[k - Stride - 4];
    int diff12 = buffer[k - Stride - 4] - buffer[k];
    int diff02 = buffer[k - Stride * 2 - 8] - buffer[k];
```



```
// if there is a diagonal gradient
if (Math.Abs(diff01) > 15 && Math.Abs(diff12) > 15)
{
    if (Math.Abs(diff02) > 30) // if there is a gradient
    {
        // increase or decrease the component depending on its value, if
        // less than 127 decrease the component value, else increase it
        buffer[k - Stride * 2 - 8] = contrastPixelComponent1(0,
            buffer[k - Stride * 2 - 8], contrastBuffer[k]);
    }
    else
    {
        // if the intensity of the component of the pixel with coordinates
        // (x-2,y-2) is greater than the one of the component of the pixel
        // with coordinates (x-1,y-1)
        if (diff01 > 0)
        {
            // turn component whiter regardless of its value
            buffer[k - Stride * 2 - 8] = contrastPixelComponent1(2,
                buffer[k - Stride * 2 - 8], contrastBuffer[k]);
        }
        else
        {
            // turn component darken regardless of its value
            buffer[k - Stride * 2 - 8] = contrastPixelComponent1(1,
                buffer[k - Stride * 2 - 8], contrastBuffer[k]);
        }
    }
}
else if (Math.Abs(diff01) > 15)
{
    if (diff01 > 0)
    {
        buffer[k - Stride * 2 - 8] = contrastPixelComponent1(2,
            buffer[k - Stride * 2 - 8], contrastBuffer[k]);
    }
    else
    {
        buffer[k - Stride * 2 - 8] = contrastPixelComponent1(1,
            buffer[k - Stride * 2 - 8], contrastBuffer[k]);
    }
}
}
```



Figure 11. The visual effect of Highlight image filter

The visual representation of the algorithm is shown in Fig. 12. The six cases next to the first `if` are the visual representation of the condition. In other words, if the condition in (5c) is met we will find ourselves in one of the six possible cases. So far for the rest of the figure, the possible cases that meet the conditions are shown next to the `if`-s and `else`-s, like it is also in the case of the second `if`, for which its condition is represented visually by two out of six possible cases.

4 Conclusions

I recommend the usage of Highlight filter rather than other filters in situations when the image contains areas of narrow text (but sure you can successfully use it also when the text in the image is wide). After applying this filter in the situation mentioned above, the success rate of OCR on the filtered image is considerably increased.

Probably the most important thing to mention regarding Highlight image filter is that it eliminates the noise in the image, more exactly, it focuses on eliminating the noise located all around the edges of characters in the text image. Beside the other actions (sharpen, contrast, highlight) of this filter, the diagonal gradient direction also contributes to removing the noise from the filtered image. The image could be also a bit blurred (not too much) and still, the OCR is improved.

In few words, Highlight filter determines outstanding OCR results on text images in which:

- Text is narrow
- Noise is present (could be around characters)
- Any other situation (e.g. lack of contrast, too much blurring)

The effect of Highlight image filter is detecting the edges and once detected it sharpens them and increases their contrast. As a result, it highlights the edges. The visual effect on characters which are present in the image would be sharpening them and increasing their contrast, creating shadows (behind them) that contrast with their color and obviously highlighting them.

In conclusion, the visual result of Highlight filter on characters is briefly described as an appropriate combination of the following visual effects, which contribute together to OCR increased performances:

- *Sharpen*
- *Contrast*
- *Highlight*

Many techniques, such as adaptive restoring of text image, have been tried, as in [5]. Image filtering has also made an improvement in important areas, such as medicine, as described in [6], but lately,



Figure 9. Possible cases while condition in (7c) is not fulfilled

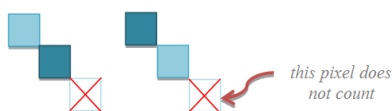


Figure 10. Possible cases for the condition in (8c) to be fulfilled

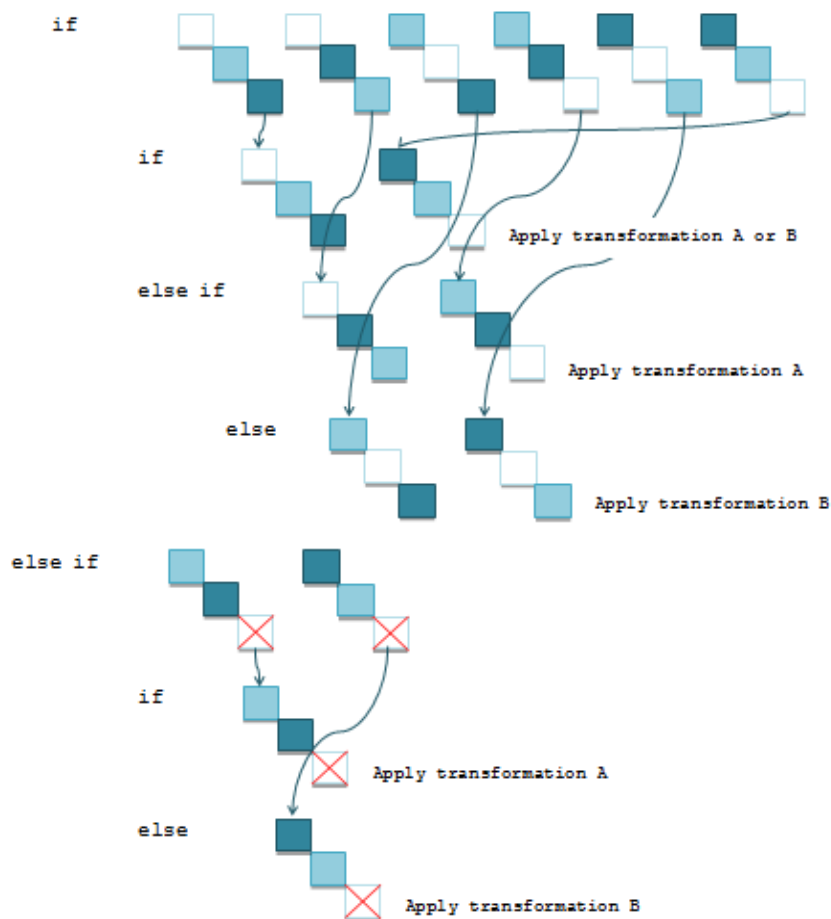


Figure 12. The visual representation of Highlight image filter algorithm for Blue component of the image

improving OCR performance using filtering has become and will be a great challenge. When developing Smart Contrast, which is a non-linear image filter, I had as a starting point the Contrast filter that was already implemented within Silicon & Software Systems Limited (S3Group) Company. Beside Smart Contrast's incontestable performances in improving OCR, this filter was actually just the triggering point for my following creation, namely Highlight image filter, as well non-linear, which overcomes the challenge of performing OCR with a very high success rate on text images.

5 References

[1] A. Rajwade, A. Rangarajan, A. Benerjee, *Automated Filter Parameter Selection Using Measures of Noisiness*, University of Florida, Gainesville, United States, 2010

[2] Tony Lindeberg, *Edge Detection and Ridge Detection with Automatic Scale Selection*, International Journal of Computer Vision, XXX, 117-156, 1998

[3] *Enhanced Filters. Sharpen*, The GIMP Help Team (Graphics by Jakob Steiner), 2001 – 2002, [Online: <http://docs.gimp.org/en/plugin-sharpen.html>]

[4] Robert Fisher, Simon Perkins, Ashley Walker, Erik Wolfart, *Laplacian / Laplacian of Gaussian*, 2003, [Online: <http://homepages.inf.ed.ac.uk/rbf/HIPR2/log.htm>]

[5] P. Stubberud¹, J. Kanai, V. Kalluri, *Adaptive image restoration of text images that contain touching or broken characters*, ¹Nevada University, Las Vegas, United States, 1995

[6] J. C. Barber¹, C. Daft², *Adaptive filtering for reduction of speckle in ultrasonic pulse-echo images*, ¹Institute of Cancer Research, Surrey, United Kingdom, ²University of Oxford, United Kingdom, 1996

SESSION
IMAGE COMPRESSION METHODS AND
TECHNOLOGIES

Chair(s)

TBA

Segmented Compression of Head Shot Photographs

Suhair Amer and Joshua Leonard

Computer Science Department, Southeast Missouri State University, Cape Girardeau,
Missouri, USA

Abstract – *Image compression is used to store a high volume of photographs in a smaller amount of space. Many methods have been developed for compressing images, all of which save space and allow for more image storage. This paper investigates segmentation that determines which parts of the face have important detail. This method then uses this distinction to improve compression without reducing picture quality.*

Keywords: segmentation, image compression, Block Truncation Coding, head shots.

1. Introduction

Law enforcement and medical applications are some of the examples utilizing image compression. Compression is usually more effective when the algorithm is tailored to a specific type of image [Michael, Goldenberg and Kimmel 2007]. With head shot photographs, we need to identify the individual while tolerating some loss in intentional specific ways to provide more powerful compression.

The first step in compressing facial photographs is to detect the outside edges to separate the face from the background [Nadernejad, Sharifzadeh and Hassanpour 2007]. Within this identified region, active and inactive regions are identified within the face [Michael, Goldenberg and Kimmel 2007]. After implementing these

algorithms a traditional compression algorithm was chosen to compress the selected blocks [Somasundaram and Sumitra 2011]. The active regions were compressed with minimal loss and the inactive regions were compressed with higher loss, but also with a much higher compression ratio. The result was an image in which the face was just as easily recognizable but with a much better overall compression ratio.

2. Project design

The first step is edge detection which is used to find major discontinuities in images. It was used to find the edges of the face as well as edges of major features (eyes, mouth, nose) [Koschan 1995]. Several methods are used for detecting edges that apply masks. For the purpose of this project, three were examined. The masks considered were the Prewitt mask [Neoh and Hazanchuk 2005], a Sobel mask [Koschan 1995], and a central differences mask [Al-Alaoui 2010].

The second step was segmenting the image. Some parts of the face are more important for recognizing a person [Bourlai, Ross and Jain 2009]. The eyes, nose, and mouth are important facial features and they clearly exhibit abrupt changes which facilitate edge detection. For this project the central differences edge detection was selected to find active and inactive regions. This was accomplished by first putting the image

through the central differences edge detector, and then examining predefined regions for the density (number of edge pixels/number of total pixels) of the edges present. If the density was over a certain threshold, the region would be considered active. Active regions could then be treated differently than inactive regions when compressing the image.

The third step was to implement and then use a traditional compression algorithm. This traditional algorithm was needed because the method being tested uses at least one underlying lossy compression method. The compression alone also serves as a point of comparison against the final algorithm. Several options were researched and considered, and the Block Truncated Coding (BTC) algorithm was chosen [Somasundaram and Sumitra 2011]. This algorithm was chosen because it was one of the simplest compression algorithms to implement while achieving reasonable compression ratios [Somasundaram and Sumitra 2011].

To summarize design stage, the image was segmented into active and inactive regions. Then the face was compressed at one level for the active regions, and at a different level for inactive regions. By using a more lossy method on inactive regions, space is saved without compromising much of the image quality. This was tested against several different levels of loss within BTC.

3. Project implementation

This project was completed using C++, with a third-party open source library ("CImg") which was used to interface with images [http://cimg.sourceforge.net/].

3.1.Segmentation

While traditional compression techniques alone provide reasonable compression rates with acceptable reconstructed image quality, this study attempts to go beyond the compression capabilities as well as image quality of these traditional methods. To achieve this, we separate regions that are important from those that are not and compress them separately.

In image compression algorithms there is a natural trade-off between compression ratio and image quality. The higher the compression ratio, the more likely an image is to be distorted. Inversely, the better quality of the reconstructed image corresponds to lower compression ratio. By separating these important and unimportant regions, compression gains can be made while saving image quality. By compressing the important and unimportant parts separately, gains can be made by using high-loss methods on the unimportant parts. This allows us to preserve the overall image quality while allowing significant gains in compression ratio.

To perform this kind of segmentation one must first find the regions that are considered important. We used an edge detection method similar to [Nadernejad, Sharifzadeh and Hassanpour 2007]. Edge detection works by finding these places where the picture changes abruptly [Bhandarkar, Zhang and Potter 1994].

For edge detection the central differences mask as shown in (Figure 1). [Neoh and Hazanchuk 2005] [Koschan 1995] [Al-Alaoui 2010] was selected,

implemented and tested.

-1/2	0	1/2
-1/2	0	1/2
-1/2	0	1/2

Figure 1: Central difference operator.

The eyes, nose, and mouth are especially important when trying to identify an individual and clearly exhibit these abrupt changes that are the inner workings of edge detection. By running an edge detection algorithm on an image, a distribution of edges can be found [Neoh and Hazanchuk 2005]. These groupings found in these distributions can be used to find these important parts of the face.

In the algorithm used in this study, the picture to be compressed is run through this edge detection. It is then divided into a discrete set of blocks. Rather than assigning blocks sizes, the algorithm is given an indication of how many columns and rows the picture should be broken into and the blocks are sized accordingly. The ratio of edges to non-edges is then calculated for each block. If the ratio exceeds a certain threshold the block is marked as being active. If this threshold is not reached, the block is marked as being inactive.

In addition a compression algorithm is implemented that when combined with the previous components will create a complete compression engine.

For this project the Block Truncation Coding (BTC) is used since it is one of the simplest algorithms as well as one that achieves acceptable compression ratios [Somasundaram and Sumitra

2011].

The BTC algorithm starts by dividing the picture into a number of discrete blocks. Each block is processed independently [Somasundaram and Sumitra 2011]. The idea is that there does not exist a great deal of variation within these blocks, so the block can be represented by only two values. Two pixel values are chosen for each block, representing the two reconstructed pixel colors. The algorithm chooses these values to best represent the block as a whole. The entire block is then assigned the high or low value using a binary 1 or 0 depending on whether the pixel is above or below the mean [Somasundaram and Sumitra 2011]. These binary assignments are stored along with the two chosen pixel colors. The larger the area of the 'Blocks' the lossier the compression, but lossier compression can sometimes be advantageous, as it produces higher compression ratios. Pseudo code for the algorithm used for segmented image compression can be found in figure 2.

1.1.Segmented Image Storage

Once the proper setup is finished this approach is fairly straight forward. First the segmentation is carried out and blocks are marked as either active or inactive. The blocks marked active are then appended to the end of the active blocks to produce a single image. The same process was repeated with the inactive blocks. Each image was then compressed using BTC, and these compressed images were written to the file along with once again the size of the picture, number of blocks, and the block numbers of the active regions.

```

void segmentedCompress(image){
    edges = detect image edges;

    listOfBlocks = split image into discrete blocks;

    for_each(block in listOfBlocks){
        if(block is active)
            add to active blocks list
        else
            inactiveBlocks.push(block);
    }

    for_each(block in active blocks list){
        append block to active image;
    }

    for_each(block in inactive blocks list){
        append block to inactive image;
    }

    compress finalActiveImage;
    compress finalInactiveImage;
}

```

Figure 2: Pseudo code for the algorithm used for segmented image compression

2. Results

The set of pictures used was a collection of 30 mugshots obtained from a mugshot database website [www.arrestcentral.com]. Each picture was processed with the same array of segmented compressions. The resulting file size and RMSE was stored for each picture. The notation used here shows the different block sizes used for the BTC compression algorithm. Compression levels are denoted by the numbers included in the labels. For example, BTC(4) means that the image was compressed with BTC with a block size of 4. The segmented compression is similar. The first number represents the block size used to compress the active regions, while the second number represents the block size used for inactive regions. For example, BTC(4/32) denotes that the image was compressed using BTC block size 4 on the active regions and block size 32 on the inactive regions.

Results of this compression are shown in Figure 3. As expected, larger BTC values yielded higher compression ratios while they also yielded worse RMSE scores (Figure 4). However, the mix of low and high values for active and inactive regions does give interesting results. For example, while the compression value rises from BTC(8/64) to BTC(16/32) (the first and second numbers being active and inactive regions respectively), the RMSE actually drops.

In an application used by law enforcement, the pictures just need to be recognizable. So the goal is to achieve the highest compression ratio possible while not leaving the picture unrecognizable. Figures 5 is an example that shows the finished photos after compression and decompression.

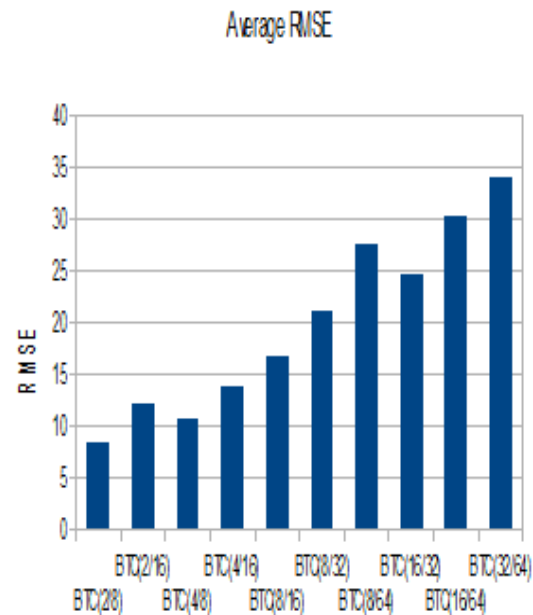


Figure 3: compression ratio.

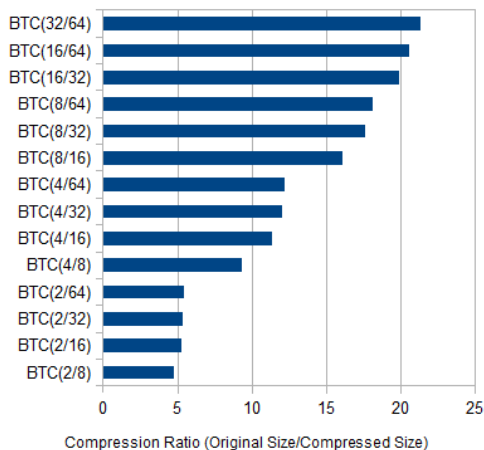


Figure 4: Average RMSE



Figure 5: Finished photos after

compression and decompression.

3. Conclusion

This paper investigated and implemented a technique for compressing head shot photographs with BTC taking advantage of segmentation.

Compression using segmentation shows that great gains were made in compressing images up until extreme BTC block sizes are reached and that the images remained recognizable. These findings prove to be useful. One of the shortcomings is that the fragments left behind by performing BTC with a block size that does not divide evenly into the size of the active and inactive region blocks. To account for this, the BTC could be reworked to have a dynamic block size or each region could be compressed by BTC individually rather than first grouping them. Future work will work on fixing these fragments and improving compression. Once the segmentation is completed, any number of compression algorithms can be applied, even using different compression algorithms for the inactive and active regions.

At the same time the RMSE (Root-mean-square deviation) measuring error varied wildly. For the BTC and Mixed BTC it increased linearly as the loss-ness.

An important part of this project to remember is that the amount of 'error' calculated between the original and compressed is not nearly as important as the ability to recognize the individual in the photograph. So error can be tolerated for the sake of compression ratio, as long as the ability to recognize

the individual is not lost.

4. Acknowledgment

This material is based upon work supported by the Grants and Research Funding Committee at Southeast Missouri State University.

5. References

- [Al-Alaoui 2010] M. Al-Alaoui. "Direct Approach to Image Edge Detection Using Differentiators"; *IEEE*, 2010.
- [Bhandarkar, Zhang and Potter 1994] S. M. Bhandarkar, Y. Zhang and W. D. Potter. "An Edge Detection Technique Using Genetic Algorithm-Based Optimization"; *Pattern Recognition*, Vol. 27, Issue 9, pp.1159-1180, 1994.
- [Bourlai, Ross and Jain 2009] T. Bourlai, A. Ross and A. Jain. "On Matching Digital Face Images Against Scanned Passport Photos"; *Proceeding of First IEEE International Conference on Biometrics, Identity and Security*, 2009.
- [Du 2006] G. Du. "Eye location method based on symmetry analysis and high-order fractal feature"; *IEEE*, Vol. 153, Issue 1, pp.11-16, 2006.
- [Koschan 1995] A. Koschan. "A Comparative Study On Color Edge Detection"; *In Proceedings of the 2nd Asian Conference on Computer Vision*, Vol. 3, pp. 574-578, Dec 1995.
- [Michael, Goldenberg and Kimmel 2007] E. Michael, R. Goldenberg and R. Kimmel. "Low Bit-Rate Compression of Facial Images"; *IEEE Transactions on Image Processing*, Vol. 16, No. 9, pp.2379-2383, 2007.
- [Nadernejad, Sharifzadeh and Hassanpour 2007] E. Nadernejad, S. Sharifzadeh and H. Hassanpour. "Edge Detection Techniques: Evaluations and Comparisons"; *Applied Mathematical Sciences*, Vol. 2, pp.507-1520, 2007.
- [Neoh and Hazanchuk 2005] H. S. Neoh and A. Hazanchuk. "Adaptive Edge Detection for Real-Time Video Processing using FPGAs"; *Altera*, 2005.
- [Somasundaram and Sumitra 2011] K. Somasundaram and P. Sumitra. "RGB & gray scale component on MPQ-BTC in image compression"; *International Journal on Computer Science and Engineering*. Vol. 3, Issue 4, pp.1462-1467, 2011.

Multiresolution image compression using non linear transformations

Ioannis Dologlou, and Stylianos Bakamidis

ATHENA–Research and Innovation Center in
Information, Communication and Knowledge Technology, Athens, Greece

Abstract - This paper presents a new non-linear algorithm for image compression operating at different resolution levels and exploiting the binary versions of the intermediate images. The decimation factor that applies between the resolution levels along with the efficient coding of the binary signals allow considerable compression rates while maintaining the image quality. Lossless compression algorithms based on arithmetic coding are used to compress the binary files that are created during the image decomposition. The method was compared experimentally against the existing standard JPEG using the well known reference image “lenna” and it was shown that for similar compression rates it is blocking artifact free, offering at the same time higher peak SNR.

Keywords: Image Compression Multiresolution non-linear

1 Introduction

Image compression is a topic that has been thoroughly investigated for many years. Many algorithms providing interesting results for all ranges of compression rates have been proposed such as the Discrete Cosine Transform (DCT) [10,11,12], the wavelet transform that offers a time-frequency representation of the image [1,2,3] and the fractals where possible self similarity within the image is identified [8,9]. Moreover multiresolution algorithms have become popular because they allow large compression rates with good performance and fast decompression at various resolutions, providing the means to the user for a quick evaluation of the content of the image. Due to the statistical properties of natural images, multiresolution representations can become very efficient. JPEG (Joint Photographic Experts Group) on the other hand is the most widely used algorithm for compressing images since it has become a standard both for black and white and colored images by ISO (International Standards Organization) and

IEC (International Electro-Technical Commission) [5]. It uses the DCT to transform the data into a set of perceptually independent features. Processing of the image is based on blocks that may become visible (blocking artifacts) when the compression rate increases.

This paper presents a new multiresolution algorithm for image compression that is based on a non linear transformation of the image at each resolution level, followed by the creation of its corresponding binary version. For image reconstruction the algorithm begins from the lowest resolution image and moves upwards by incorporating the information within the binaries at the different resolution levels. Both analysis and synthesis parts of the methodology are simple and fast to implement. Since the approach is not block based it does not present any blocking artifacts, not even for high compression rates.

2 Image Analysis-Decomposition

This section presents the decomposition algorithm of the image in order to obtain components that are easier to compress at various resolution levels. First the image I is converted to zero mean by subtracting its global mean m . The zero mean image is decomposed into two separate images A_I and B_I , where A_I represents the absolute value of I and B_I represents the sign of I . By employing the matlab notation for pixel by pixel multiplication of two images it is straightforward to show that,

$$I = m + A_I \cdot B_I \quad (1)$$

The above expression maybe interpreted as the decomposition of the original image I into a product of its envelope image A_I and its corresponding sign image B_I . At this stage the multiresolution scheme is introduced by considering the smoothed version of the envelope image A_I instead of the original one. Smoothing of A_I is achieved using a two dimensional low pass filter, thus creating A_{IS} . Depending on the cutoff frequency of the filter, A_{IS} can be decimated

accordingly leading to a lower resolution level. The decimated version of AIS stands for the new original image at the lower resolution level, namely I_{i+1} , and it is further processed by repeating the same steps. The binary image BI_i at all resolution levels, is compressed to CBI_i without loss using arithmetic coding algorithms [6,7]. The block diagram of

Figure 1 shows graphically the i -th step of the decomposition algorithm.

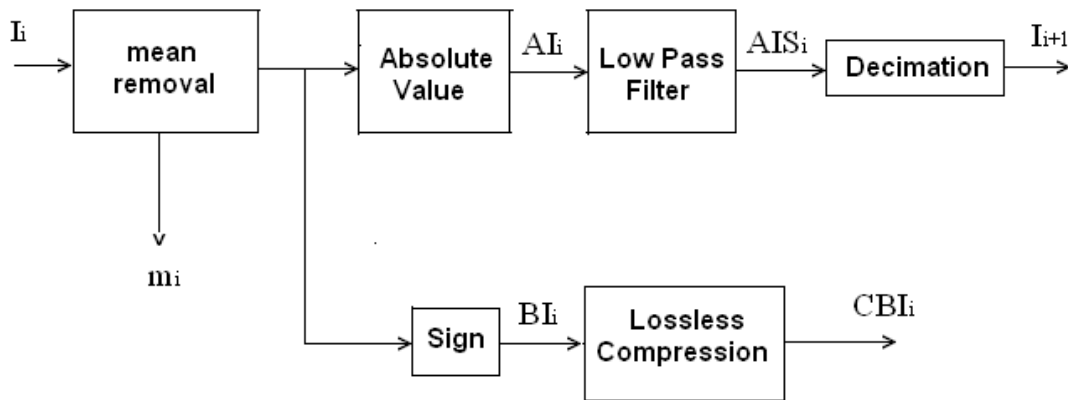


Figure 1. Block diagram of the i -th resolution level of the Analysis-Decomposition Algorithm.

3 Image Synthesis-Reconstruction

This section presents the reconstruction algorithm from the components that have been created and stored during the analysis stage. Assuming $i+1$ is the current resolution level, the i level image is reconstructed using CBI_i along with m_i and AIS_i that comes from the interpolation of I_{i+1} i.e. the image from the previous resolution level. First CBI_i is decompressed to obtain BI_i and it is then combined with AIS_i and m_i to compute I_i according to

$$I_i = m_i + AIS_i * BI_i \quad (2)$$

The new image I_i is further processed by repeating the same steps. The block diagram of Figure 2 shows graphically the reconstruction algorithm.

Successful reconstruction requires the storage of all the different CBI_i 's together with the means m_i 's and the lowest resolution level image I_n , where n stands for the number of steps of the algorithm (resolution levels).

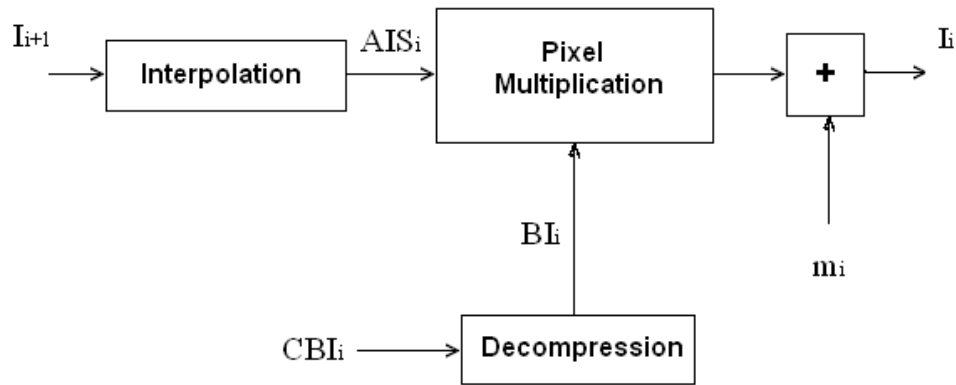


Figure 2. Block diagram of the i -th resolution level of the Synthesis-Reconstruction Algorithm.

4 Experimentation-Testing

The new image compression algorithm offers two parameters to determine the desirable compression ratio. The first parameter refers to the number of resolution levels n and the second parameter refers to the decimation factor d that is used at each resolution level. It is advisable to use a balanced combination of these two parameters to obtain best results for a given compression rate. On one hand large decimation factors provide fast data shrinking from one resolution level to another, implying though considerable loss of information during envelope smoothing (small cutoff frequency of the low pass filter). On the other hand a small decimation factor, although it preserves well the quality of the smoothed envelopes, it creates large intermediate binary images. In our tests a variety of different combinations was used and it was shown experimentally that for a given image size and decimation factor it is always better to use the maximum possible resolution levels n in order for the lowest resolution level image I_n to have minimum dimensions.

The new proposed algorithm was tested against the existing standard JPEG. Comparisons are based on the reference image “lenna” with dimensions 512×512 . The results shown below refer to 20.8 compression ratio. The compression ratio was chosen as high to enable a better appreciation of the differences of the two methods as artifacts become more visible. The decimation factor for the new method was set to $d=0.69$ and the number of resolution levels n was set to 8. The size of the lowest resolution image I_8 is 27×27 . For the same compression ratio JPEG gives a smaller peak SNR equal to 32.66 db while the new method gives 32.93 db. Figure 3 shows the compressed image of the new algorithm and Figure 4 depicts the compressed image of JPEG at the same compression ratio 20.8. It can be seen that the new technique has no blocking artifacts while JPEG has (around the shoulder area).



Figure 3. The original “lenna” (left) and the compressed image using the new method (right).



Figure 4. The original “lenna” (left) and the compressed image using JPEG (right).

In addition Figures 3.1 and 4.1 show the zoomed version of the face of the images where it can be noted that the new method (Fig. 3.1) preserves better the edges than JPEG (Fig 4.1). It is also interesting to focus on the eyes of the two images and note the improved performance of the new method. The middle image (Fig. 5) depicts the zoomed version of the original “lenna”.

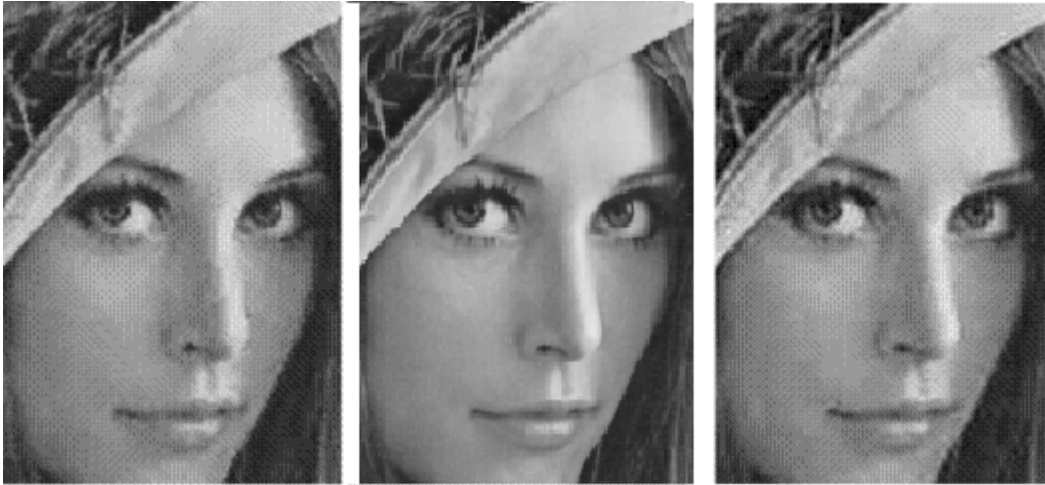


Figure 3.1
Zoomed face of the
new method

Figure 5.
Zoomed face of the
original "lenna"

Figure 4.1
Zoomed face of JPEG

5 Conclusion

A new multiresolution algorithm for image compression using a non linear transformation of the images at different resolution levels was presented. The method was compared against the existing standard JPEG and was shown to perform better in terms of peak SNR and blocking artifacts.

Future research will focus on one hand on the optimal selection of the decimation factor versus the number of resolution levels and on the other hand on the use of more efficient lossless compression techniques of the binary signals.

6 References

- [1] J.S. Walker, "A primer on wavelets and their scientific applications", 2nd ed. Chapman & Hall/CRC, 2008.
- [2] A. Zandi, D. Allen, E. Schwartz, and M. Boliek, "CREW: compression with reversible embedded wavelets," Proc. IEEE Data Compression Conf., 1995.
- [3] S. Mallat, "A wavelet tour of signal processing", Academic Press, 1998.
- [4] B.Babb, S. Becke, and F. Moore, "Evolving optimized matched forward and inverse transform pairs via genetic algorithms," Proc. 48th IEEE Int. Midwest Symp. Circuits and Systems, 2005.
- [5] G.Wallace, "The JPEG still picture compression standard", Communications of the ACM 34(4): 30-44, 1991.
- [6] P.G. Howard, J.S. Vitter, "Arithmetic coding for data compression", Proc. IEEE, 82(6): 857-865, June 1994.
- [7] Sayood, Khalid, "Introduction to Data Compression", Third Edition. Morgan Kaufmann Publishers, pp. 560-569, 2006.
- [8] E. Arnaud, Jacquin, "Image Coding Based on a Fractal Theory of Iterated Contractive Image Transformations", IEEE Transactions on Image Processing, 1(1), 1992.
- [9] J. Kominek, "Advances in fractal compression for multimedia applications", Journal on Multimedia Systems, Volume 5 Issue 4, July 1997, Pages 255 - 270
- [10] Wen-Hsiung Chen, C. Smith, S. Fralick, "A Fast Computational Algorithm for the Discrete Cosine Transform". IEEE Transactions on Communications 25 (9): 1004-1009, 1977.
- [11] B. Andrew, Watson, "Image Compression Using the Discrete Cosine Transform", Mathematica Journal, 4(1), 81-88, 1994
- [12] Prabhakar, Telagarapu, V. Jagan Naveen, A. Lakshmi Prasanthi, G.Vijaya Santhi, "Image Compression Using DCT and Wavelet Transformations", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 3, September, 2011

Symmetrical Compression of Facial Photographs

Suhair Amer and Joshua Leonard

Department of Computer Science, Southeast Missouri State University, Cape Girardeau, MO, USA

Abstract – *Image compression allows for a higher volume of photographs to be stored in the smaller amount of space. Many methods have been developed for compressing images, all of which save space. This paper introduces a method for combining a well-known compression technique with a technique that takes advantage of the symmetry of the human face.*

Keywords: symmetry, image compression, Block Truncation Coding, head shots

1. Introduction

Image compression is an important component of image processing of digital photographs. The finite available memory, makes it very important to find compression techniques that would save space. For example, improved compression techniques are especially important for applications that are used by law enforcement, where there are an enormous number of mugshots taken each year. Law enforcement needs to be able to store all of these pictures at a high enough quality to be recognizable while at the same time not using up all of their limited resources.

This project focused on compression that would be specifically tailored towards these kinds of photographs. Since we know the type of the compressed image, the chosen compression technique can be more effective [Michael, Goldenberg and Kimmel 2007]. Since photographs such as mugshots must identify the individual to some extent, some loss can be tolerated in intentional specific ways to provide more powerful compression.

For this project, several methods of compressing facial photographs were developed and implemented before symmetry were applied and tested.

First an algorithm was used to detect the edges in the photographs [Nadernejad, Sharifzadeh and Hassanpour 2007]. The next step was to detect active and inactive regions within the face [Michael, Goldenberg and Kimmel 2007]. Once these algorithms had been implemented, a traditional compression algorithm was used for compressing the image before applying symmetry and after for comparison reasons [Somasundaram and Sumitra 2011]. Similar to [Du 2006] edge detection was used to find a vertical line that corresponds to the face's line of symmetry. This took advantage of the natural symmetry of the face and only

half of the image was stored. During decompression, the missing half was reconstructed using the half stored.

2. Project design

The project was broken into several components/stages. In the first stage, several edge detection techniques were investigated to detect/separate the face from the background. Edge detection is used to find major discontinuities in images, it was used to find the edges of the face as well as the edges of major features (eyes, mouth, nose) [Koschan 1995]. Three methods were considered for edge detection, all of which applied masks over the image. The following masks were considered : the Prewitt mask [Neoh and Hazanchuk 2005] , a Sobel mask [Koschan 1995], and a central differences mask [Al-Alaoui 2010]. Using the masks, a pixels value is determined by adding weighted values that surround the pixel. In the Prewitt and Sobel masks, the algorithm is weighted towards either side of the pixel and each diagonal, and then again for the top and bottom [Koschan 1995][Neoh and Hazanchuk 2005]. The difference between the two is that the Sobel mask weights the direct horizontal or vertical pixels more than the diagonals. In the central differences mask, the diagonals are disregarded completely [Al-Alaoui 2010]. After implementing and testing the different masks, the central differences mask was chosen for its simplicity and speed.

In the second stage, we developed a system that compresses facial photographs and that takes advantage of the symmetry of the human face. While traditional compression algorithms alone provide reasonable compression rates with acceptable reconstructed image quality, this study aimed to take the compression ratio even higher. We attempt to take advantage of the natural properties of the human face, namely the natural symmetry of features [Du 2006]. The eyes lie across equidistant from one another, and the mouth and nose are both obviously symmetric. Taking advantage of this symmetry, this study finds a line of symmetry within the face region, and uses it to reduce image size. If it can be shown that the portions on either side of the line are sufficiently symmetric. Accordingly, we decided to store half of the facial region to represent both sides.

The third step was to implement a traditional compression algorithm. Several options were researched and considered, and the Block Truncated Coding (BTC) algorithm was chosen [Somasundaram and Sumitra 2011]. It was chosen because it was simple to implement and still achieved reasonable compression ratios [Somasundaram and Sumitra 2011].

3. Project implementation

This project was completed using C++, with a third-party open source library ("CImg") used to interface with images

[<http://cimg.sourceforge.net/>].

3.1. Symmetrical Compression

In theory, symmetry could be used to achieve huge gains in compression ratio. There are a number of ways to detect symmetry within a facial photograph [Kiryati and Gofman 1997][Du 2006]. One method relied on a symmetry line starting from one side and incrementally making its way to the other side and testing the strength of symmetry at each point [Kiryati and Gofman 1997]. Although this approach is very simple, there is an inherent trade-off for time and accuracy. The increment at which the line would be moved at each iteration directly influences this trade-off. If the line is moved too little the processing takes too long, but if the line is moved too much the likelihood of finding the strongest point of symmetry is greatly reduced.

Because we wanted a symmetry algorithm that is fast and easy to implement, we decided to detect the symmetry line by first identifying the features that cause edges in the face regions such as the eyes, nose, and mouth. After running the edge-detection algorithm on a copy of the image, the symmetry of edges is used to find the central point of symmetry in the face. For example, horizontally, we can identify the edges of both eyes and have a central point between both as point one. As we continue on, we can identify

the outer edges of both sides of the nose and identify center point as point two. We and identify the outer edges of the mouth and identify the center point as point three. We can connect a vertical line between points one, two and three and consider it as our symmetry line. This algorithm would both process images very quickly and have no restrictions on the accuracy level provided. This is because because it is not calculated by moving in set intervals.

We chose edge based symmetry detection. In theory it is both fast and accurate and therefore advantageous over the more processing-intensive and potentially inaccurate method.

3.2. Edge based symmetry detection

Edge Based Symmetry detection takes advantage of some simple principles. The human face has several specific features that stand out from the face as a whole. The eyes, nose, and mouth in particular are thought of as being very symmetrical across the face [Du 2006]. This works well, as some features also provide large discontinuities (the cheeks, forehead, chin, and other unremarkable regions of skin) from the background. These discontinuities make them an ideal candidate for edge-detection which relies wholly on the abrupt change of pixel value to locate an edge.

To find this symmetry, we scanned the image for edges from both the left and the right at the same vertical position. When an edge is found from both directions, the point that is horizontally in the center of them is marked as that vertical position's point of symmetry. This process is repeated for the entire picture and these points of symmetry are averaged together to produce an approximate overall symmetry line. Vertical positions were excluded from the calculation if the points found by both sides were in the same horizontal location (noise) to prevent them from heavily influencing the overall result.

Next, some statistical principles were applied. The standard deviation for the set of suspected symmetry point was calculated. The points that lied outside the standard deviation, were excluded from the calculation. This prevented outliers from heavily influencing the outcome of the overall symmetry line. The Pseudocode in figure 1 describes this process.

```

unsigned findLineOfSymmetry(image)
{
edges = detect edges in image;

for(each y coordinate){
leftEdge = first edge from left side at y coordinate;
rightEdge = first edge from right side at y coordinate;

if(leftEdge != rightEdge){
increment number of points;
add the midway point of leftEdge and rightEdge to the list of
points;
}
}

averageX = the average of all the x coordinates in the list of points;
standardDeviation = find standard deviation of the list of points;

for_each(point in the list of points){
if(point is further than standardDeviation from avg){
remove point from the list of points ;
}
}

finalLineOfSymmetry = the average of the list of points ;
return finalLineOfSymmetry;
}

```

Figure 1: Pseudocode for the algorithm used for Edge Based Symmetry detection

4. Results

The implemented algorithm was tested against 30 mugshots. Multiple block sizes were used for the BTC compression, shown by the number appended onto the BTC prefix. The average compression ratio was calculated (Figure 2).

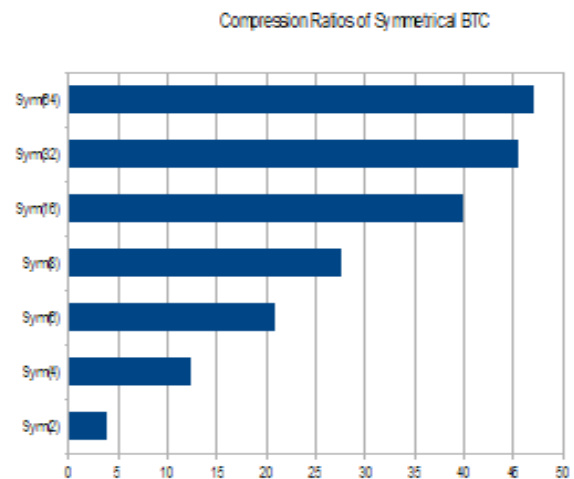


Figure 2: Compression ratio calculated by dividing the size of the original image by the size of the compressed image across different block sizes (2, 4, 6, 8, 16, 32, 64)

In addition, Figure 3 clearly shows that although using symmetry greatly increases the compression ratio, it produces a reconstructed image that is very different than the one that was

compressed.

In most cases this can be overlooked when it comes to person being recognizable. The high error rate can result from the offset of the reconstructed side of the face compared to the original.



Figure 3: reconstructed images with different block sizes.

5. Conclusion

In this paper we developed a system to compress head shot photographs with BTC taking advantage of symmetry.

Compression with symmetry was shown to make great gains in compressing file sizes. However, certain kinds of facial photographs result in poor quality. The reasons for the poor results include the inability to compensate for individual pictures that are not symmetrical. This results from having asymmetrical faces. Asymmetrical facial tattoos and piercings are an obvious problem, as well as asymmetrical hair styles. In addition, this algorithm does not account for a person's face being rotated or tilted to one side or the other. When the head is rotated to one side, part of the face can be lost because of the lack of symmetry in the original photograph. When the face is tilted, the line of symmetry is inaccurate as the algorithm currently only finds symmetry in a completely straight vertical lines. Multiple angles could be examined to try to find the exact tilt of a persons face.

At the same time the RMSE (Root-mean-square deviation) that measures error varied wildly. For the BTC and Mixed BTC it increased linearly as the lossiness. As for the symetric compression, the RMSE spiked extremely high at all compression levels. This is attributed to the fact that it is checked against the original image and although it may look like the original, one half of the image was not included in the original.

An important part of this project is to remember that the amount of error

calculated between the original and compressed images is not nearly as important as the ability to recognize the individual in the photograph. So error can be tolerated for higher compression ratios, as long as the ability to recognize the individual is not lost. The compression algorithm was found to be generally successful. When the symmetrical compression was run, the results were extremely varied. In the cases where it worked well, we achieved higher compression and still have a recognizable face. In cases where it did not work well, the faces and their features were distorted past the point of recognition. This stemmed from the inability of the chosen method to correct itself especially with faces that were tilted one way or the other. Certain photos produced other problems such as having hair extending down into faces. Some improvements could be made such as correcting tilt and perhaps requiring an image to reach a certain threshold of symmetry before deciding to allow it to be compressed using the symmetry.

6. Acknowledgment

This material is based upon work supported by the Grants and Research Funding Committee at Southeast Missouri State University.

7. References.

[Al-Alaoui 2010] M. Al-Alaoui. "Direct

Approach to Image EdgeDetection Using Differentiators"; *IEEE*, 2010.

[Du 2006] G. Du. "Eye location method based on symmetry analysis and high-order fractal feature"; *IEEE*, Vol. 153, Issue 1, pp.11-16, 2006.

[Kiryati and Gofman 1997] N. Kiryati and Y. Gofman. "Detecting Symmetry in Grey Level Images: The Global Optimization Approach"; *International Journal of Computer Vision*. Vol. 29, Issue 1, pp.29-45, 1997.

[Koschan 1995] A. Koschan. "A Comparative Study On Color Edge Detection"; *In Proceedings of the 2nd Asian Conference on Computer Vision*, Vol. 3, pp. 574-578, Dec 1995.

[Michael, Goldenberg and Kimmel 2007] E. Michael, R. Goldenberg and R. Kimmel. "Low Bit-Rate Compression of Facial Images"; *IEEE Transactions on Image Processing*, Vol. 16, No. 9, pp.2379-2383, 2007.

[Nadernejad, Sharifzadeh and Hassanpour 2007] E. Nadernejad, S. Sharifzadeh and H. Hassanpour. "Edge Detection Techniques: Evaluations and Comparisons"; *Applied Mathematical Sciences*, Vol. 2, pp.507-1520, 2007.

[Neoh and Hazanchuk 2005] H. S. Neoh and A. Hazanchuk. "Adaptive Edge Detection for Real-Time Video Processing using FPGAs"; *Altera*, 2005.

[Somasundaram and Sumitra 2011] K. Somasundaram and P. Sumitra. "RGB & gray scale component on MPQ-BTC in image compression"; *International Journal on Computer Science and Engineering*. Vol. 3, Issue 4, pp.1462-1467, 2011.

Image Compression of Colored Facial Photographs: Segmentation and Symmetry-Combined

Suhair Amer and Joshua Leonard

Computer Science Department, Southeast Missouri State University, Cape Girardeau, Missouri, USA

Abstract – *Compression of facial photographs raises research challenges because such images are used widely in agencies such as police and law-enforcement, schools and universities, states. It also requires efficient storage. In this paper human frontal facial images with a plain background are considered. Combined segmentation and symmetry techniques are examined which produced better compression rates when compared to using either method individually. Although image quality is reduced, the segmentation preserves the ability to recognize the faces.*

Keywords: segmentation, symmetry, image compression, Block Truncation Coding, head shots.

1. Introduction

This project focused on compression of facial photographs. Choosing the compression technique was straightforward because we knew what kind of image we are processing [Michael, Goldenberg and Kimmel 2007]. Since the project aim is to be able to identify the individual after decompressing the image, some loss can be tolerated in intentional specific ways to provide more powerful compression.

To compress the image, the image had to go through several stages.

First an algorithm was used to detect edges in photographs [Nadernejad, Sharifzadeh and Hassanpour 2007]. This means that the face region was identified or separated from the background. The next stage was to detect active and inactive regions within the face region [Michael, Goldenberg and Kimmel 2007]. The active regions were compressed with minimal loss algorithm and the inactive regions were compressed with much more loss, but also with a much higher compression ratio. Then a compression algorithm was chosen to compress the image [Somasundaram and Sumitra 2011]. The result was an image in which the face was just as easily recognizable but with a much better overall compression ratio.

The final stage was to find a vertical line that corresponds to the face's line of symmetry [Du 2006]. Taking advantage of the natural symmetry of the face, only half of the image was stored, and in the decompression stage, the missing half was reconstructed using the half stored.

2. Project design

For the first stage, the face was separated from the background by identify the edges of the face using a central differences mask [Al-Alaoui 2010]. In the masks, a pixel's value is determined by adding the weighted values that surround the pixel. In the central

differences mask, the diagonals are disregarded completely. In general, it was chosen for its simplicity and speed.

The second stage was segmenting the image. Some parts (regions) of the face are considered more important than others especially when it comes to identifying the person after decompressing the image [Bourlai, Ross and Jain 2009]. The eyes, nose, and mouth are important and clearly exhibit these abrupt changes that are the inner workings of edge detection. Using the central differences edge mask detection, we find active and inactive regions. This is accomplished by first putting the image through this central differences edge detector, and then examining predefined regions for the density of the edges present. If it was over a certain threshold, the region would be considered active. Active regions could then be treated differently than inactive regions while compressing the image.

The third stage was to run the image through a Block Truncated Coding (BTC) algorithm [Somasundaram and Sumitra 2011]. It is a simple algorithm to use and still achieves reasonable compression ratios.

By using a more lossy method on inactive regions, space is saved without compromising much of the image quality. This method was implemented with the ability to utilize several different levels of loss within BTC.

The final stage was to test and utilize symmetry. While traditional compression algorithms may provide reasonable compression rates with acceptable reconstructed image quality, this study aimed to take the compression

ratio even higher. We attempt to take advantage of the natural properties of the human face, namely the natural symmetry of features [Du 2006]. The eyes lie across equidistant from one another, and the mouth and nose are both obviously symmetric. A line of symmetry was found, then one half was stored and used to reconstruct the whole face in the decompression stage.

3. Project implementation

This project was implemented using C++, with a third-party open source library ("CImg") that was used to interface with images [<http://cimg.sourceforge.net/>].

3.1. Segmented compression

In the segmentation stage, we separate regions that are important from those that are not, and compress them separately and differently.

Since there is a tradeoff between compression rate and image quality, we have segmented or separated the face region into important and unimportant regions. This will achieve compression gains while saving image quality. We will be using high-loss methods on the unimportant parts which will result in low quality parts of the image but these regions are considered not significant.

To perform segmentation we need to identify discontinuities which can be found through edge detection. Edge detection works by finding these places where the picture changes abruptly [Bhandarkar, Zhang and Potter 1994]. The central differences mask [Neoh and Hazanchuk 2005] [Koschan 1995] [Al-Alaoui 2010] was selected.

To identify the important features of the face that are the eyes, nose, and mouth, we run an edge detection algorithm on an image, and then a distribution of edges can be found [Neoh and Hazanchuk 2005] . These groupings found in these distributions can be used to find these important parts of the face. The image is divided into a discrete set of blocks. The ratio of edges to non-edges is then calculated for each block. If the ratio exceeds a certain threshold the block is marked as being active. If this threshold is not reached the block is marked as being inactive.

The BTC algorithm starts by dividing the picture into a number of discrete blocks. Each block is processed independently and then represented by only two values [Somasundaram and Sumitra 2011]. These two pixel values are chosen for each block, representing the two reconstructed pixel colors. The algorithm chooses these values to best represent the block as a whole. The entire block is then assigned the high or low value using a binary 1 or 0 depending on whether the pixel is above or below the mean [Somasundaram and Sumitra 2011]. These binary assignments are stored along with the two chosen pixel colors.

Once all blocks are processed and information is stored to a file, the segmentation stage is complete.

3.2.Symmetrical Compression

The next stage is to symmetry. For the purpose of this project we choose to detect symmetry using an edge-detection algorithm. It is run on a copy of the image, and then the symmetry of edges is used to find the central point of

symmetry in the face. This algorithm would both process images very quickly and, because it is not calculated by moving in set intervals, have no restrictions on its level of accuracy.

Symmetry detection based on edges takes advantage of some simple principles. The human face has the eyes, nose, and mouth that are thought of as being very symmetrical across the face [Du 2006]. They also provide large discontinuities from the background on which they are placed (the cheeks, forehead, chin, and other unremarkable regions of skin).

Therefore, the edges are scanned from both left and right at the same vertical position. When an edge is found from both directions, the point that is horizontally in the center of both is marked as that vertical position's point of symmetry. This process is repeated for the entire picture and these points of symmetry are averaged together to produce an approximate overall symmetry line. Vertical positions were excluded from the calculation if the points found by both sides were in the same horizontal location to prevent unimportant part of the image from heavily influencing the overall result.

3.3.Combined Segmented and symmetrical compression

The goal of this study was to store these pictures in a very small format. While a traditional compression algorithm can be powerful by itself, we are more interested with the combination of normal compression with segmented and symmetry detection.

To improve the compression ratios, we try to reduce the amount of data that the chosen compression has to store. When a line of symmetry is formed we discard one half of the image. When reconstructed, the idea is that the faces natural symmetry will give a recognizable face even if it is simply half of the face flipped to complete itself.

This approach at compression begins with the edge based symmetry detection. Once the face's line of symmetry has been found this value is used to effectively discard one side of the image. The remaining side was compressed by itself and the data for this compression was stored to a file, along with an indication that the image has been cut at its symmetry point (Fig. 4).

To reconstruct the image the reverse approach is taken. First the compressed data is decoded, yielding one half of the face. This half is then copied, flipped, and appended onto itself providing an artificial second half.

Both methods presented here have shown gains in compression size while maintaining a reasonable amount of quality. To further increase the compression ratios we combine these methods together.

The symmetric compression leaves half an image to be compressed using any method that would work on a normal image. So, after the symmetric processing is done the image can be segmented and compressed as was done with the other photographs.

Once the individual pieces have been put

into place, the implementation of the combination is fairly straight forward.

To summarize, the face regions are first identified and then are processed by the symmetry detection algorithm. Then half of the face region is processed and passed to the segmentation stage where the active and inactive regions are identified. The data is then stored to a file.

To reconstruct, the opposite approach is taken. The active blocks and inactive blocks are decompressed yielding half of the image. The image is then reflected to yield the final result.

4. Results

The developed system was tested against 30 facial photographs, and the average compression ratio as well as the RMSE data was calculated.

Our goal was to achieve the highest compression ratio possible while not leaving the picture unrecognizable. Figures 1 shows an example of the result of using the combined Symmetric and Segmented compression.

It should be noted that although using symmetry greatly increases the compression ratio, it produces a reconstructed image that is very different from the original. In some cases this can be overlooked if the person is recognizable. In one image, for example, although the calculated RMSE was very high the person could still be easily recognized.

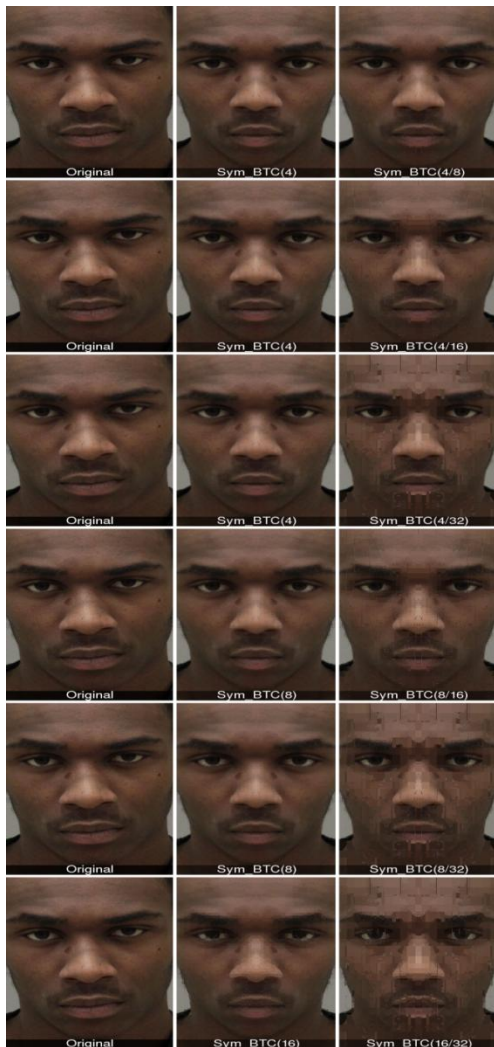


Figure 1: Reconstructed images using Combined Symmetric and Segmented compression.

Figure 2 shows the Compression ratio calculated by dividing the size of the original picture by the size of the compressed picture. Figure 3 shows the Average calculated RMSE. In general, the combined segmented and symmetrical algorithm outperformed using each one individually while still producing recognizable images.

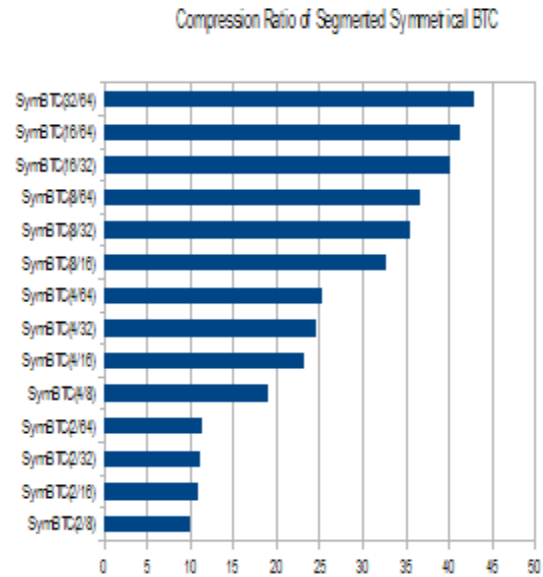


Figure 2: Compression ratio calculated by dividing the size of the original picture by the size of the compressed picture.

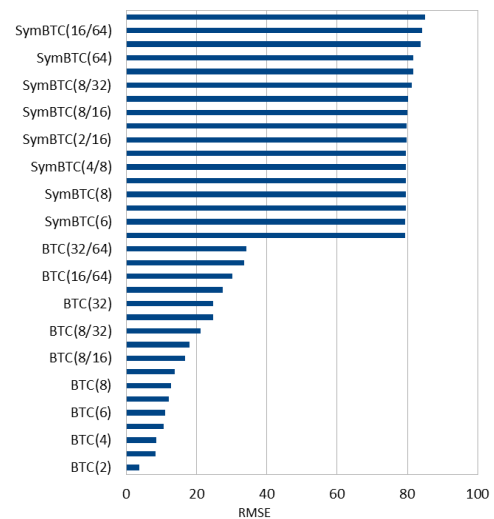


Figure 3 : Average RMSE

Compression ratios started around 6:1 for the least lossy algorithm that used only BTC compression. As other methods such as symmetry and segmentation were added, these methods produced a wide range of results. For example, the compression using symmetry achieved a ratio of over 45:1. RMSE scores varied widely and were much higher for images compressed with symmetry.

5. Conclusion

This paper experimented with techniques to compress head shot photographs with BTC taking advantage of segmentation and symmetry.

When both symmetry and segmentation were combined, better compression rates were reached than either method individually. Although image quality is reduced, the segmentation preserves the ability to recognize the faces. This method is a feasible way to compress images for use in large databases of facial photographs.

At the same time the RMSE (Root-mean-square deviation) that measures error varied wildly. It increased linearly while using segmentation. When adding symmetric compression, the RMSE spiked extremely high at all compression levels. This is because we are checking the reconstructed image (consisting of one half of the face with the other half replica of the second half) against the original image.

An important part of this project is to remember that the amount of error calculated between the original and compressed is not nearly as important as the ability to recognize the individual in

the photograph. So error can be tolerated for the sake of compression ratio, as long as we are still able to recognize the individual. All of the compression algorithms were found to be generally successful. The segmented compression produced much improved compression ratios, and although it caused the inactive regions of some pictures to look odd, the oddness did not prevent the face from being recognized. When the symmetrical compression was run, the results were extremely varied. In cases where it worked well, huge compression gains were found while leaving a perfectly recognizable face. When it did not work well, the faces and their features were distorted past the point of recognition. This problem stemmed from the inability of the chosen method to correct for faces that were tilted one way or the other. Certain photos produced other problems, such as images of individuals whose hair extended down into their faces. Some improvements could be made such as correcting for tilt and perhaps requiring an image to reach a certain threshold of symmetry before deciding to allow it to be compressed using the symmetry.

6. Acknowledgment

This material is based upon work supported by the Grants and Research Funding Committee at Southeast Missouri State University.

7. References

- [Al-Alaoui 2010] M. Al-Alaoui. "Direct Approach to Image EdgeDetection Using Differentiators"; *IEEE*, 2010.
- [Bhandarkar, Zhang and Potter 1994] S. M. Bhandarkar, Y. Zhang and W. D. Potter. "An Edge Detection Technique Using Genetic

- Algorithm-Based Optimization”; *Pattern Recognition*, Vol. 27, Issue 9, pp.1159-1180, 1994.
- [Bourlai, Ross and Jain 2009] T. Bourlai, A. Ross and A. Jain. “On Matching Digital Face Images Against Scanned Passport Photos”; *Proceeding of First IEEE International Conference on Biometrics, Identity and Security*, 2009.
- [Du 2006] G. Du. “Eye location method based on symmetry analysis and high-order fractal feature”; *IEEE*, Vol. 153, Issue 1, pp.11-16, 2006.
- [Koschan 1995] A. Koschan. “A Comparative Study On Color Edge Detection”; *In Proceedings of the 2nd Asian Conference on Computer Vision*, Vol. 3, pp. 574-578, Dec 1995.
- [Michael, Goldenberg and Kimmel 2007] E. Michael, R. Goldenberg and R. Kimmel. “Low Bit-Rate Compression of Facial Images”; *IEEE Transactions on Image Processing*, Vol. 16, No. 9, pp.2379-2383, 2007.
- [Nadernejad, Sharifzadeh and Hassanpour 2007] E. Nadernejad, S. Sharifzadeh and H. Hassanpour. “Edge Detection Techniques: Evaluations and Comparisons”; *Applied Mathematical Sciences*, Vol. 2, pp.507–1520, 2007.
- [Neoh and Hazanchuk 2005] H. S. Neoh and A. Hazanchuk. “Adaptive Edge Detection for Real-Time Video Processing using FPGAs”; *Altera*, 2005.
- [Somasundaram and Sumitra 2011] K. Somasundaram and P. Sumitra. “RGB & gray scale component on MPQ-BTC in image compression”; *International Journal on Computer Science and Engineering*. Vol. 3, Issue 4, pp.1462-1467, 2011.

SESSION
POSTERS AND SHORT PAPERS

Chair(s)

TBA

Motion Analysis of Horse-Rider for Coaching System

Jae-Neung Lee, Myung-Won Lee, Yeong-Hyeon Byeon, Keun-Chang Kwak¹

¹Department of Control and Instrumentation Engineering, Chosun University, Gwangju, Korea

Abstract - This paper construct a database of the national representative level of a professional horse-rider by wearing a motion-capture suit attached with 16 inertial sensors under an inertial sensor-based wireless network environment, then make a visual comparative analysis on the values of all motion features (elbow angle, knee angle, knee-elbow distance, backbone angle and hip position) classified depending on Warm-blood and obtained by various methods of calculating Euclidean distance, the second cosine, maximum and minimum values, and made a comparative analysis depending on motion features of a horse-rider by using MVN studio software. In the study, the experimental results confirmed the validity of the proposed method of obtaining the motion feature database of a horse-rider in the wireless sensor network environment and making an analytical system.

Keywords: Inertial sensor, horse riding, wireless sensor, motion analysis

1 Introduction

People are doing exercises to keep a good body shape. Especially, everybody knows that obesity is good neither for appearance nor for health. Horse-riding is a good sport of keeping good health and body line. It is possible to analyze and properly coach the postures of a horse-rider by the analysis of the so-called horse-riding motions under the wireless sensor network environment. Particularly, horse-riding is an exercise with a special trait that a horse-rider and a horse alive should be joined together. It can be helpful for a horse-rider to build up physical health and spiritual growth. In addition, the horse-riding is a physical exercise of the whole body helpful to improve body's balance and flexibility for general physical developments. However, the horse-riding is not effective if correct body postures are not learned and maintained properly in the course of the exercise. Therefore, in order to make the most effective achievement within the shortest period of time, it is necessary to get a coaching session to check what is wrong with the motions of a horse-rider. Recently a lot of studies have been made by using inertial sensors at the wireless sensor network environment to carry out the most effective coaching session of the horse-riding sport. Yujin [1] suggested Upper Body Motion Tracking with Inertial Sensors. Lijun [2] analyzed A Practical Calibration Method on MEMS Gyroscope. Wei [3] suggested calibration of low-precision MEMS inertial sensor. Cao [4]

performed 3D dynamics analysis of a golf full swing by fusion inertial sensors and vision data. Jung [5] analyzed smart shoes. Chan [6] proposed a virtual reality dance training system using motion capture technology. Frosio [7] analyzed automatic calibration of MEMS accelerometers. However, a variety of inertial sensor-based studies have been made so far, but no study has been made about an analysis on the motions of a horse-rider in a wireless network environment. Thus, in this paper, a database is constructed by collecting the motions of a professional horse-rider wearing a motion-capture suit under the inertial sensor-based wireless network environment. At this time, the representative type of horse, named Warm-blood, was selected for this study. Actual horse-riding sessions were made to analyze the postures of a professional horse-rider by measuring and calculating all the motion feature values of elbow angle, knee angle, knee-elbow distance, backbone angle and hip y-axis position. The results of the experiments made at MVN Studio with MATLAB confirmed the validity of the method of analyzing the motions of a horse-rider in a wireless sensor network environment suggested in this study, in which all the motion data of a horse-rider were collected in the wireless sensor network environment to make a comparative analysis on all the horse-riding postures carefully.

2 Inertial sensor-based motion analysis

This chapter describes a method of constructing an inertial sensor-based wireless network environment and a motion database. All the motion data of a horse-rider are received to a computer through the MVN motion capture system constructed with inertial sensors. Then, all the data are compared by using respective calculation methods. Differently from an optical sensor-based motion capture system, the MVN motion capture system can capture the entire body motions wirelessly without using a camera. In addition, the MVN motion capture system is portable for convenient indoor-outdoor uses.

In order to collect data, this study used a subject, a national representative level of a professional horse-rider whose height is 164cm with her foot size of 235mm. The inertial sensor-based 3D motion capture suit made by Xsens, to collect data, subsequently riding on the horses, Warm-blood whose height is 150~173cm. Fig. 1 shows the construction of horse rider's motion by Warm-blood. The period of time taken for measurement of one file was about 1~2 minutes and 15 data were collected depending on footpace types. There are 4 horse footpace types such as walk, trot, canter and gallop. A horse usually goes as far as 130m for a minute, approximately

8km for an hour at a walk. It usually moves as far as 220m for a minute, approximately 13km for an hour at a rising trot, one specific type of trots. It generally moves as far as 350m for a minute, approximately 21km for an hour at a canter. It moves as far as 1000m for a minute, approximately 60km or even 72km for an hour at a gallop. The horses used in the experiments of this study were made to move at the two footpace types, a rising trot and a canter. The measurement frame rate was 100 frames per second. If body postures are not properly learned and maintained by a horse-rider, the horse-riding exercise may bring not a positive, but a negative effect. Therefore, it is necessary to make an analysis on the horse-riding postures to clearly check what is wrong with the motions of the horse-rider. The analysis methods [8] are presented to make a comparative analysis on 5 motion features (elbow angle, knee angle, elbow-knee distance, backbone angle, hip position).



Fig. 1 Construction of motion database by Warm-blood

3 Experimental results

In the experiments of this study, data were collected by Warm-blood at two footpace types (rising trot and canter) 15 times, so that 4 cycles of data were extracted for comparative analysis. It was confirmed that there was differences in horse-rider's postures through different values of respective angles and distances. The right elbow and knee are marked in red dotted lines. Fig. 2 shows that repeating 200 frames were extracted out of approximately 10000 frames to demonstrate 4 cycles (2.5 seconds). As shown in the Fig. 2, the elbow and knee angles remain at the range of about 130~150 degrees and about 130-170 degrees, respectively. The elbow and knee distances stay at the range of 18~27cm and 15~18cm, respectively. The backbone angle remains at the range of 170~177 degrees. The Hip positions moves within the range of 30~37cm. Fig. 3 shows that repeating 300 frames were extracted out of approximately 10000 frames to demonstrate 4 cycles (3 seconds). As shown in Fig. 3, the elbow and knee angles remain at the range of about 130~160 degrees and 130~140 degrees, respectively. The elbow and knee distances stay at the range of 25~29cm and 15~18cm, respectively. The backbone angle remains at the range of 170~177 degrees. The Hip positions moves within the range of 30~37cm.

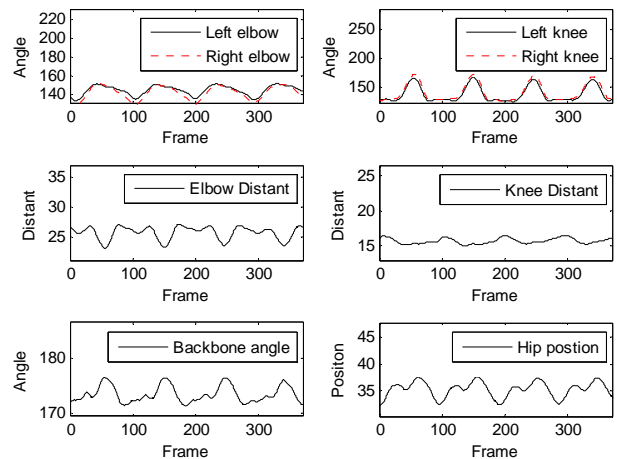


Fig. 2 Feature values of Warm-blood at a trot

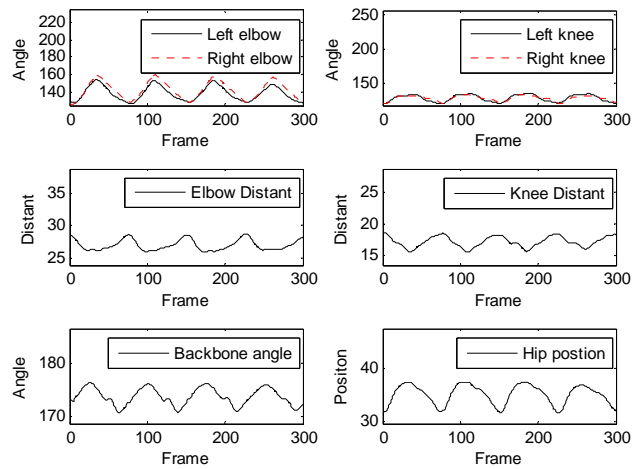


Fig. 3 Feature values of Warm-blood at a canter

Fig 4 below illustrates the numerical comparison of maximum and minimum feature values of Warm-blood at a canter. A visible difference is revealed in the elbow angles. A similar difference is noticed in knee angles as shown in the elbow angles. As described above, no significant difference was made in the backbone angles and in the hip positions because the horse-rider should keep her backbone at its perpendicularity and her hip at the same position regardless of the type of horses.

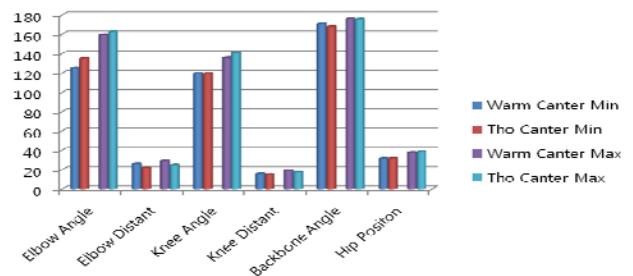


Fig. 4 Comparison of maximum and minimum feature values at a canter

4 Conclusions

This paper suggested a method of using a motion database of a professional horse-rider wearing a suit constructed with wireless networks consisting of 16 inertial sensors and then extracting the respective motion features (elbow angle, knee angle, backbone angle, hip position, knee-elbow distance) through various calculation methods such as Euclidean distance, the second cosine, maximum and minimum values, depending on Warm-blood and footpace types (trot and canter). MVN studio software was used to make a comparative analysis on the horse-rider's motion features depending on the footpace types. As a result, a significant difference was noticed in the motion feature values obtained depending on different horse and footpace types. Therefore, in order to effectively make real-time coaching sessions for different horse-riding footpace types, it is necessary to construct a motion feature database in relation to footpace types and accordingly make a suitable analysis and coach on horse-riding motions.

5 Acknowledgements

This work was supported by the Ministry of Knowledge Economy (MKE), Rep. of Korea, under the IT R&D program supervised by the KOREA Evaluation Institute of Industrial Technology (KEIT) (10041059)

6 References

- [1] J. Yujin, K. Donghoon, K. Jinwook, "Upper Body Motion Tracking With Inertial Sensors," *Proc. IEEE International Conference on Robotics and Biomimetics*, pp. 1746-1751, 2010.
- [2] S. Lijun, Q. Yongyuan, "A Practical Calibration Method on MEMS Gyroscope". *Piezoelectrics & Acousto-optics*, Vol. 32, No. 3, pp. 372-374, 2010.
- [3] R. Wei, Z. Tao, Z. haiyun, W. Leigang, "A Research on Calibration of Low-Precision MEMS Inertial Sensors," *IEEE Trans. on Control and Decision Conference*, pp. 3243-3247, 2013.
- [4] N. Cao, S. Young, K. Dang, "3D Dynamics Analysis of a Golf Full Swing by Fusing Inertial Sensor and Vision data," *International Conference on Control, Automation and Systems*, pp. 1300-1303, 2013.
- [5] P. G. Jung, G. Lim, K. Kong, "A Mobile Motion Capture System Based on Inertial Sensors and Smart Shoes," *IEEE Trans. on Robotics and Automation*, pp. 692-697, 2013.
- [6] J. C. P. Chan, H. Leung, J. K. T. Tang, and T. Komura, "A virtual reality dance training system using motion capture technology," *IEEE Trans. on Learning Technologies*, Vol. 4, No. 2, pp. 187-195, 2011.
- [7] I. Frosio, F. Pedersini, and N. A. Borghese, "Auto calibration of MEMS accelerometers," *IEEE Trans. on Instrumentation and Measurement*, Vol. 58, No. 6, pp. 2034-2041, 2009.
- [8] M. W. Lee, K. C. Kwak, "3D Motion analysis of national rider athletes by riding types in horse simulator", *Third International Conference on Innovative Computing Technology (INTECH 2013)*, August, London, UK, pp.12-16, 2013.

Adaptive Fruit and Vegetable Products Recognition based on multi-feature fusion

Huawei Tao¹, Cairong Zou^{1,2}, and Li Zhao¹

¹Key Laboratory of Underwater Acoustic signal Processing of Ministry of Education, Southeast University, Nanjing, Jiangsu, China

²Guangzhou University, Guangzhou, Guangzhou, China

Abstract - Automatic recognition of fruits and vegetable products via computer vision is still a difficult task due to complexity of environment. This paper develops a robust fruit and vegetable products recognition method, which integrates color and texture feature based on Harmonic adaptive feature fusion (HAFF) algorithm proposed by this paper. HAFF firstly use exp function to improve AWC, then compute different weights mean to form a new set of weights. We use HAFF to fuse HSV color histogram, BIC color histogram and completed local binary pattern texture histogram as a hybrid feature; finally, nearest neighbor algorithm is used for multi-class classification. Matlab Simulation result shows that: fusing color features and texture feature by HAFF can acquire the best performance.

Keywords: Products recognition; Feature fusion; Computer Vision; Color histogram; texture histogram;

1 Introduction

Accurate identification of different kinds of vegetable and fruit products is an important quality for cashier in supermarkets, so the supermarkets must invest time, money training qualified cashiers; through the use of barcodes can help cashier to point out different kinds of products price, the cashier must prepackage products and paste barcodes, which consume a lot of manpower and material resources and limit the choice of customers.

Recently, the scale using image to adaptively identify products attracts public attention. Bolle et al. [1-2] proposed "Veggie Vision" which was the first supermarket products recognition system. However, "Veggie Vision" is sensitive to the illumination, so it have high demand for illumination; Anderson Rocha et al.[3] combine several individual features from the state-of-the-art to assess how they interact to improve the overall accuracy of the system; S. Arivazhagan [4] proposed that using H, S color histogram feature and Co-occurrence texture feature to identify fruit and vegetable products; Anderson Rocha et al. [5] proposed a new fusion method based on binary problems to classify different kinds of fruit and vegetable products; Fabio Augusto Faria et al. [6] presented a novel framework for classifier fusion aiming at supporting the automatic recognition of fruits and vegetables

in a supermarket environment; Yudong Zhang et al. [7] propose a novel classification method based on a multi-class kernel support vector machine (kSVM) to identify fruit and vegetable products.

Above mentioned techniques can obtain good recognition effect in some cases, however, there are still many questions when those techniques were used in real environment. Firstly, most of texture operators are sensitive to the illumination; secondly, although some of above methods can achieve good accurate rate, fusion methods need a lot of computation;

In this paper, we construct a database which captures products pictures under six different illuminations to estimate algorithm. We proposed a new feature fusion method called Harmonic adaptive feature fusion (HAFF) algorithm, which fuse color feature, texture feature as hybrid feature, finally, nearest neighbor algorithm is used for multi-class classification. Matlab Simulation shows that using HAFF to fuse features can acquire the best result.

The rest of the paper is organized as follow: Section 2 shows the proposed fusion algorithm; Section 3 shows the experimental protocol; section 4 shows experimental result and discussion.

2 Features Fusion Method

2.1 Multimodal biometric fusion

Traditional multimodal biometrics fusion method [8] contains three stages: matching score, normalization, fusion. In matching score stage, the matching score can be acquired by calculating similarity between two feature vectors. In the normalization stage, the matching score is normalization.

In the fusion stage, according to the normalization score, the features have been fused. The classical normalization algorithms contain: Simple-Sum (SS), Min-Score (MIS), Max-Score (MAS).

Simple-Sum (SS):

$$f_i = \sum_{m=1}^M n_i^m \quad (1)$$

Where n_i^m represents the normalized score for matcher m ($m = 1, 2, \dots, M$ where M is the number of matchers)

applied to user i ($i = 1, 2, \dots, I$, where I is the number of individuals in the database). f_i is the fused score for user i .

2.2 Adaptively Weighted Cue

Adaptively Weighted Cue (AWC) is proposed by Paul Brasnett et al. [9], which is used in object tracking in video sequences. It uses the reciprocal of smallest of the distance measurements as cue weight. The AWC can be defined as:

$$\varepsilon = \frac{1}{D_{l,\min}^2}, l = 1, 2, \dots, L \quad (2)$$

$$\hat{\varepsilon} = \frac{\varepsilon}{\sum_{l=1}^L \varepsilon_l}, l = 1, 2, \dots, L \quad (3)$$

Where $D_{l,\min}$ is the smallest value of the distance measure for cue l , $\hat{\varepsilon}$ is normalized weight.

2.3 Harmonic Adaptive Feature Fusion Algorithm

Adaptively Weighted Cue can do well in object tracking. However, features with different dimensions are prone to have different order of magnitude similarity, which will make individual ε far larger than others. In order to improve AWC, we firstly use function \exp to balance difference of different feature weight. Motivated by 2.1 and improved AWC, we propose a new fusion algorithm called Harmonic adaptive feature fusion (HAFF) algorithm.

Step 1: Distance vector $dist^i, (i = 1, 2, \dots, N)$ can be got through calculating the feature i similarity between test sample and the training database. Get the minimum value $dist_{\min}^i$ of $dist^i, (i = 1, 2, \dots, N)$;

Step 2: Get the reciprocal of $dist_{\min}^i = \min(dist^i), (i = 1, 2, \dots, N)$ and normalize it. Obtain the weighting vector w ;

$$w(i) = \frac{\frac{1}{dist_{\min}^i}}{\frac{1}{dist_{\min}^1} + \dots + \frac{1}{dist_{\min}^N}}, (i = 1, 2, \dots, N) \quad (4)$$

Step 3: Compute the exponential function of w and normalize it. Obtain the exponential weighting vector w' ;

$$w'(i) = \frac{\exp(w(i))}{\exp(w(1)) + \dots + \exp(w(N))}, (i = 1, 2, \dots, N) \quad (5)$$

Step 4: Calculate the mean of w and w' and normalize it. Obtain the harmonic adaptive weighting vector w'' ;

$$w''(i) = \frac{w(i) + w'(i)}{\sum_{i=1}^N w(i) + \sum_{i=1}^N w'(i)}, (i = 1, 2, \dots, N) \quad (6)$$

Step 5: Get the fusion vector of distance vector based on the feature fusion method

$$dist = \sum_{i=1}^N w''(i) * dist^i \quad (7)$$

3 Experimental Protocols

3.1 Multimodal biometric fusion

Fig.1 depicts fruit and vegetable product image acquisition device. The vertical distance between the camera and pan is 32 cm. The camera is the three hundred thousand megapixel camera module. The fruit and vegetable product images are saved in JPEG format; the image size is 640pi*480pi. It takes us a month to construct fruits and vegetable products data set, and the data set contains 13 kinds of products under 6 kinds of illumination. Figure 2 show the product under 6 kinds of illumination. The database contains: Apple, Banana, broccoli, cucumber, Dragon fruit, grape, Green vegetable, peach, Dangshan pear, potato, tomato, kiwi fruit, lettuce. Figure 3 show all kinds of fruit and vegetable products.

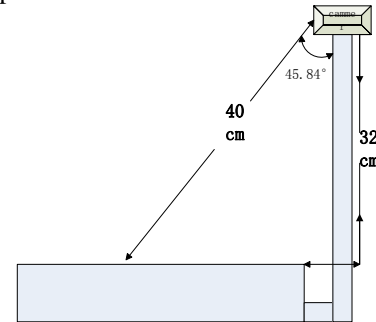


Fig1 Fruit and vegetable image acquisition device



Fig2 Fruit and vegetable image under 6 kinds of illumination

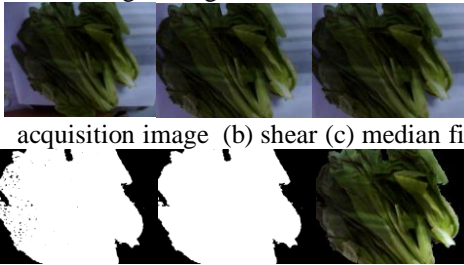




Fig3 fruite and vegetalbe image

Illumination shadows, occlusions will have a great effect on the images, and those cases will affect the system recognition accuracy. In order to improve recognition accuracy, image preprocessing method should be done. The preprocessing method is shown as below:

- Step 1:** image taking by camera often contains a portion. So unnecessary interference background is cut;
- Step 2:** the image is processed through 3*3 Median filter after cutting;
- Step 3:** Convert the RGB color image into Lab color space. Use the K-mean method to segment the image;
- Step 4:** Fill the image;
- Step 5:** Remove image background;



(a) acquisition image (b) shear (c) median filter

(d) segmentation (e) hole filling (f) remove background.

Fig 4 image processing

3.2 Image descriptor

Extracting features is the key to the success of fruit and vegetable product recognition. In this paper, we choose HSV color histogram, Border/interior pixel classification (BIC) [10] and CLBP [11] as image descriptors.

4 Experimental Protocols

Figure 4 show products classification algorithm flow sheet. We use the χ^2 statistics function as the dissimilarity between two features. The χ^2 statistics function can be calculated as equation (8). We use nearest neighbor algorithm to indentify the class of fruit and vegetable.

$$d_{\chi^2}(H, K) = \sum_{i=1}^B \frac{(h_i - k_i)^2}{h_i + k_i} \quad (8)$$

Our database contains 13 kinds of product image; Dragon fruit and grape have 60 images (each illumination has 10 images), other products have 120 images (each illumination has 20 images). All the pictures are acquired with different number of products. We choose images acquired from one illumination as training database and other 5 kinds of illumination images as testing database.

Table 1 show the simulation result. BIC+CLBP+HSV represent multi-feature fusion without fusion method; (BIC+LBP+HSV)_SS represent multi-feature fusion with Simple-Sum multi-modal biometrics fusion algorithm; (BIC+LBP+HSV)_FAFF represent multi-feature fusion with Harmonic adaptive feature fusion algorithm; According to the table 1, The following observation can be made. Firstly, CLBP can achieve better classification accuracy rate than Unser and TextA; this result show that CLBP is robust to the illumination, and Unser and TextA are sensitive to the illumination. Secondly, we can see that (BIC+LBP+HSV)_HAFH can achieve better accuracy rate than other algorithm. Compare with multi-feature without fusion algorithm, the (BIC+LBP+HSV)_HAFH mean rate can achieve best recognition.

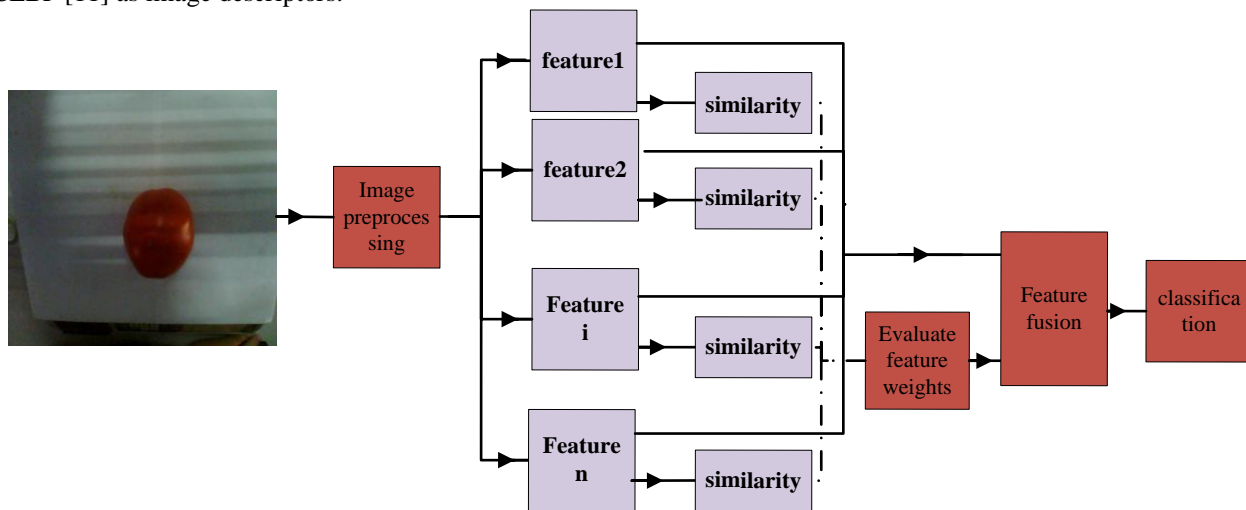


Fig4: Products Classification Algorithm Flowsheet

Table1 experimental result on database

	Rate/%						mean
	Illumination 1	Illumination 2	Illumination 3	Illumination 4	Illumination 5	Illumination 6	
Unser[5]	19.00	37.08	36.42	38.92	37.92	40.58	34.99
TextA[1]	20.25	25.00	26.33	31.25	32.08	30.25	27.53
Algorithm CLBP	44.67	53.75	55.17	56.25	60.50	60.83	55.20
BIC+CLBP+HSV	62.75	79.50	75.25	76.00	79.00	78.92	75.24
(BIC+CLBP+HSV)_SS	61.67	77.67	72.58	74.67	76.67	78.58	73.46
(BIC+CLBP+HSV)_ HAFF	63.25	79.75	77.50	75.92	79.58	78.67	75.78

5 Conclusion

Automatic visual recognition of products is not as unattainable a goal. Related methods have been proposed by researcher. In this paper we propose a new fusion method which fuse color features and texture feature to indentify kinds of fruit and vegetable products. The result shows that our method can acquire better result than before study.

Acknowledgements

This research project was founded in part by National Natural Science Foundation of China (No. 61273266, No. 61375028).

6 References

- [1] Bolle, R. M., J. H. Connell, N. Haas, R. Mohan and G. Taubin. Produce recognition system, United States Patent: 5,546,475. 1996.8.13.
- [2] Bolle, R. M., J. H. Connell, N. Haas, R. Mohan and G. Taubin. "Veggievision:A produce recognition system". in Proceedings 3rd IEEE Workshop on Applications of Computer Vision, 1996, pp.244-251.
- [3] Rocha, A., D. C. Hauagge, J. Wainer and S. Goldenstein. "Automatic produce classification from images using color, texture and appearance cues". Presented at SIBGRAP'08. XXI Brazilian Symposium on Computer Graphics and Image Processing, 2008, pp.3-10.
- [4] Arivazhagan, S., R. N. Shebiah, S. S. Nidhyandhan and L. Ganesan. "Fruit recognition using color and texture features" Journal of Emerging Trends in Computing and Information Sciences, vol.1, no.2, pp.90-94. Oct 2010.
- [5] Rocha, A., D. C. Hauagge, J. Wainer and S. Goldenstein. "Automatic fruit and vegetable classification from images" Computers and Electronics in Agriculture, vol.70, no.1, pp.96-104, Jan 2010.
- [6] Faria, F. A., J. A. d. Santos, A. Rocha and R. d. S. Torres. "Automatic Classifier Fusion for Produce Recognition". presented at 25th SIBGRAP Conference on Patterns and Images, 2012,pp.252-259.
- [7] Zhang, Y. and L. Wu. "Classification of fruits using computer vision and a multiclass support vector machine" sensors, vol.12, no.9, pp. 12489-12505, 2012.
- [8] Snelick R, Mink U U A, Indovina M, et al." Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems".IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, no. 3, pp. 450-455, Mar 2005.
- [9] Brasnett, P., L. Mihaylova, D. Bull and N. Canagarajah (2007). "Sequential Monte Carlo tracking by fusing multiple cues in video sequences." Image and Vision Computing, vol. 25, no.8, pp. 1217-1227, Aug 2007.
- [10] Stehling, R. O., M. A. Nascimento and A. X. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In proceedings of the eleventh international conference on Information and knowledge management, 2002, pp.102-109.
- [11] Z.Guo, L.Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," IEEE trans. Image Process., vol.19, no.6, pp.1657-1663, June 2010.

Real-time 3D Hand Pointing using Depth Image

Sung-II Joo¹, Sun-Hee Weon¹, and Hyung-II Choi¹

¹Department Global Media, Soongsil University, Seoul, Republic of Korea

Abstract - This paper proposes a real-time 3D hand mouse method using depth information, for the purpose of providing users with a more intuitive interface that controls smart devices through gestures. The process for creating the hand mouse method interface is as follows. First, the hand region is detected using OpenNI based on information from the depth image, and the central point of the hand is calculated. The central point of the hand is used as the standard point for gesture recognition and tracking, and is therefore referred to as tracking point. Using this tracking point, we detect the direction of the palm of the hand, and calculate the point of intersection between a straight line extended in the tracked direction of the palm and the plane in the 3D virtual space, then convert this into the coordinates of the actual device that we wish to control.

Keywords: 3D hand pointing, Depth image, Gesture interface, Hand mouse

1 Introduction

Gesture interfaces have long been a topic of research in the field of computer vision. The gesture interface methods that are currently available in commercialized smart devices, however, mostly consist of technologies that codify the shape or movement of the hands. Although such methods can have the effect of enhancing the rate of recognition, they require a preceding stage of gesture learning, and the disadvantage is that the user is required to remember a large number of gestures.

To develop the hand mouse, S. W. Kang [1] and C. H. Hwang [2] used skin color information to detect the hand. The problem with this method is that it is sensitive to changes in lighting condition. I. B. Jeon [3] used Kinect to detect the hand region using pre-defined hand size information from the depth image and then counted the number of fingers to process the mouse click event. This method is feasible only when there is no obstructing object between the camera and the location of the hand, and it is difficult to overcome the problem caused by the fact that the size of the hand region changes depending on the location of the user. To solve these problems, this paper proposes a new hand mouse method which uses depth images to eliminate the effects of lighting conditions and which uses a plane approximation of the palm of the hand in order to recognize the direction of the palm, rather than using the endpoint of the finger which affects detection performance.

2 3D Hand Pointing

Interfaces using existing hand mouse methods require the user to make a relatively large movement in order to move the mouse to a point that is distant from its current position. By contrast, the 3D pointing method proposed in this paper makes it possible to control large spaces more easily by simply using information on the movements of the palm's direction and the hand's location.

$$ax + by + c = z \quad (1)$$

$$A = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, X = \begin{bmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{bmatrix}, Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \quad (2)$$

$$XA = Z \quad (3)$$

$$A = (X^T X)^{-1} X^T Z \quad (4)$$

Equation (1) - (4) expresses for the plane using the Least Squared Method, the purpose of which is to detect the plane of the hand region. When there is an equation for the plane as in equation (1), the expression is given as a determinant as shown in equation (2) in order to utilize multiple items of data. Expressed as a determinant, formula (1) can be represented as equation (3), and the final equation of the plane can be obtained by calculating equation (4), which has been converted by pseudo inverse.

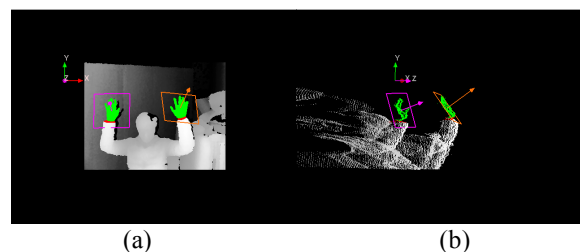


Fig. 1. Results of detection the palm's plane and direction using the plane approximation

[Fig. 1] is the result of using the equation explained above to detect the palm's plane and direction (normal vector). The direction of the palm has been expressed as the plane's normal vector based on the average location of the interest pixels, and as a result, as shown in the figure, it becomes

possible to achieve a stable detection performance even when the palm is not aligned flatly to the camera.

As explained above, it is possible to obtain the direction vector even by using the plane's equation. However, if the plane's normal vector is defined by the direction vector for each frame, the detection results can become unstable due to the various noises that are present in the depth image. To address this problem, therefore, a weighted value is given as follows so that the direction can be gradually detected.

$$E_t = [S_t(x) \ S_t(y) \ S_t(z)] - l[a \ b \ -1] \quad (5)$$

$$E_t' = E_t \alpha + E_{t-1}'(1-\alpha) \quad (6)$$

S_t is the central coordinate of interest pixels, and E_t indicates the direction vector's endpoint that was detected in the t^{th} frame when S_t is taken as the basis. Equation (5) is the formula for obtaining E_t , a and b refers to the parameter obtained in equation (4), and l refers to the length of the direction vector. Equation (6) is used to perform interpolation using the previous direction vector and the current direction vector, and the interpolation is made by α . The final direction vector can therefore be expressed as $\vec{n} = (E_t' - S_t) / \|E_t' - S_t\|$, with S_t being the original point.

3 Experiment Result

In this study, the test was performed on a depth image with a size of 320*240, obtained through the input device Kinect, manufactured by Microsoft.

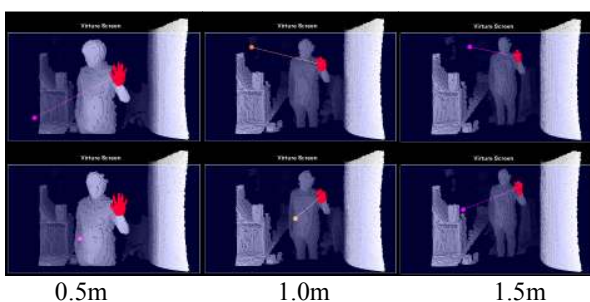


Fig. 2. Results of palm plane and direction detection at varying distances

[Fig. 2] shows results of performing palm plane approximation and the detection of the direction vector according to the method proposed in this study. The test was performed with the distance between the user and the camera set variously within a range of 0.5m~1.5m. This test confirmed that the proposed method is capable of palm plane detection even when the size of the hand changes, or in other words, even when the user is moving freely within a certain range of distance from the controlled device.

[Fig. 3] shows the results of applying 3D hand pointing in order to control the device using the detected direction vector

and shows the resulting detection of the intersection point. The dark blue quadrangle is the virtual screen, and the red circle and green circle mark the intersection between the virtual screen and the direction of the palm.

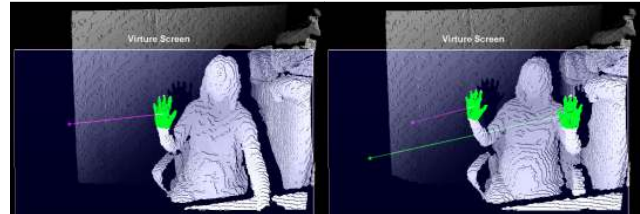


Fig. 3. Results of detecting the 3D hand pointing intersection for virtual screen control

4 Conclusions

This paper proposed a 3D hand pointing method using depth images to controlling devices. In this method, the tracking point of the hand region that was detected from the depth image is used to perform plane approximation and detect the plane of the palm and then the vector of the plane's direction is obtained in order to realize the gesture interface which uses the hand mouse method for virtual screen control. This method requires no prior definitions or learning processes, and can perform well even in response to changes in lighting or the environment.

Acknowledgements

This work was supported by the Seoul R&BD Program(SS110013) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning(2013R1A1A2012012).

5 References

- [1] S. W. Kang, C. J. Kim and W. Sohn, "Developing User-friendly Hand Mouse Interface via Gesture Recognition", Proceedings of Conference on The Korean Society of Broadcast Engineers, pp.129-132, 2009.
- [2] C. H. Hwang, H. H. Lee, Y. J. Jeon and J. I. Kim, "Design and Implementation of Human Interface Technology Realization based on Hand Mouse", Journal of The Research Institute of Industrial Technology Development, Vol.26, pp.75-81, 2012.
- [3] I. B. Jeon and B. H. Nam, "Implementation of Hand Mouse Based on Depth Sensor of the Kinect", Proceedings of 43th Conference on KIEE, pp.1674-1675, 2012.
- [4] S. H. Park, S. J. Yu, J. R. Kim, S. J. Kim and S. Y. Lee, "3D hand tracking using Kalman filter in depth space", EURASIP Journal on Advances in Signal Processing, Vol.2012, No. 1, pp.1-18, 2012.

Method for Correction of Lenses Distortion in Stereo Vision

Osmando Pereira Junior¹, Joceli Mayer²

¹Federal Institut of Triângulo Mineiro / Campus Paracatu, MG - Brazil

²Digital Signal Processing Laboratory / Federal University of Santa Catarina - Florianopolis - SC - Brazil

osmandoj@gmail.com, joceli.mayer@lpds.ufsc.br

Abstract—*Lenses distortion is one of the main factors that limits the accuracy of stereo vision system reconstruction. We propose a new method for correction of the lenses distortion by applying compensation to each region of an image. Our method splits the image into smaller regions and compensates for each region for a fixed lenses model order. When compared to the conventional method, which models the entire image with only one model, our approach provides considerably better compensation and reduce the depth error as shown in the experiments with synthetic data.*

Keywords: Lens Distortion, Camera Calibration, Stereo Vision, Image Processing

1. Introduction

Stereo Vision is the process of recovery of three-dimensional information of a scene, or an object of scene, from the analysis of two bidimensional images, by using an appropriate camera model [1], [2]. The cameras allow for a rich representation of the scene when compared to other types of sensors, such as laser and sonar, being used more and more in applications for mobile robotics and assistance driving, such as object and obstacle detection and localization [3]–[5]. Stereo vision is also used in remote sensing and Metrology [6], and is composed of three main steps: camera calibration, pixel correspondence and 3-D reconstruction [2], [7].

Calibration is the process of estimating intrinsic and extrinsic camera parameters and is an essential task to achieve reconstruction accuracy. The lenses distortions decreases the calibration precision and the reconstruction accuracy, thus a proper compensation method for lenses is required. In this paper we analyse the effect of the model order of the lenses and propose a novel local compensation method. Results are shown to contrast to traditional methods.

2. Stereo Vision Model

A point in the three-dimensional space is represented by $\mathbf{M} = [X, Y, Z]^T$. The projection of \mathbf{M} on the image frame is represented by \mathbf{m} . The index ‘ w ’ (\mathbf{v}_w) is used to denote a point represented on the 3D coordinate system R_w of world, with measuring unit mm , whereas the index ‘ i ’ (\mathbf{v}_i) is used to denote a point represented on the image 2D coordinate system R_i , and is measured in pixels. $\tilde{\mathbf{v}}$ is

used to represent a point in homogeneous coordinates, where $\tilde{\mathbf{v}} = [\mathbf{v}^T \ 1]^T$ and $\hat{\mathbf{v}}$ to stand for a real multiple of $\tilde{\mathbf{v}} = a\tilde{\mathbf{v}}$, where a corresponds to a scale factor. Indexes ‘ l ’ and ‘ r ’ are used to represent points on images I_l on the left and I_r on the right, respectively.

The projection of \mathbf{M}_w on the image frame R_i is given by equations (1) and (2) [2], [7].

$$\hat{\mathbf{m}}_{i_{[pixel]}} = \begin{bmatrix} \eta_u & 0 & 0 \\ 0 & \eta_v & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & u_0 \\ 0 & -f & v_0 \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{R}|\mathbf{t}] \tilde{\mathbf{M}}_w \quad (1)$$

$$\tilde{\mathbf{m}}_i = \frac{\hat{\mathbf{m}}_i}{\hat{m}_{i_3}} \quad (2)$$

where η_u and η_v represent the pixel dimensions on image sensor directions u and v , respectively; f is the focal length of the camera; (u_0, v_0) is the principal point coordinates; \mathbf{R} is the rotation matrix of the world coordinate system R_w for camera coordinate system R_c , whose determination is presented in [2]; \mathbf{t} is the translation of the camera center in relation to \mathbf{O}_w , the origin of R_w , represented in R_c ($\mathbf{t} = -\mathbf{R}\mathbf{C}_w$).

Once both intrinsic and extrinsic camera parameters, and the homologous pixels pairs (\mathbf{m}_{l_w} , \mathbf{m}_{r_w}) are known, projections of \mathbf{M}_w on images I_l and I_r , respectively, it is possible to reconstruct \mathbf{M}_w by applying equation (3).

$$\begin{aligned} \mathbf{M}_w &= \lambda_1 (\mathbf{m}_{l_w} - \mathbf{C}_{l_w}) + \mathbf{C}_{l_w} \\ \mathbf{M}_w &= \lambda_2 (\mathbf{m}_{r_w} - \mathbf{C}_{r_w}) + \mathbf{C}_{r_w} \end{aligned} \quad (3)$$

where λ_1 and λ_2 are scale factors calculated by (4), and \mathbf{m}_w , the projection of \mathbf{M}_w in the image frame, represented in world system coordinates R_w , is determined by (5).

$$\begin{aligned} &[\lambda_1 \quad \lambda_2] \left(\begin{bmatrix} \mathbf{m}_{l_w}^T \\ -\mathbf{m}_{r_w}^T \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{l_w}^T \\ -\mathbf{C}_{r_w}^T \end{bmatrix} \right) + \\ &+ [1 \quad 1] \begin{bmatrix} \mathbf{C}_{l_w}^T \\ -\mathbf{C}_{r_w}^T \end{bmatrix} = \mathbf{0}^T \end{aligned} \quad (4)$$

$$\mathbf{m}_w = \mathbf{R}^T \begin{bmatrix} \mu_u & 0 & 0 \\ 0 & \mu_v & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -u_0 \\ 0 & -1 & v_0 \\ 0 & 0 & f \end{bmatrix} \tilde{\mathbf{m}}_i + \mathbf{C}_w \quad (5)$$

where $\mu = \frac{1}{\eta}$; the principal point coordinates, u_0 and v_0 , as well as the coordinates of \mathbf{m}_i , are measured using pixel units; and focal length f is measured in mm .

3. Proposed Local Lenses Compensation

A traditional lenses model consists of 3 components as follows: radial, decentering and thin prism distorton, given by equations (6), (7) and (8), respectively [8], [9].

$$\delta_r(\mathbf{m}) = (\check{k}_1 \|\mathbf{m}\|^2 + \check{k}_2 \|\mathbf{m}\|^4 + \dots + \check{k}_n \|\mathbf{m}\|^{2n}) \mathbf{m} \quad (6)$$

$$\delta_d(\mathbf{m}) = \begin{bmatrix} 2x^2 + \|\mathbf{m}\|^2 & 2xy \\ 2xy & 2y^2 + \|\mathbf{m}\|^2 \end{bmatrix} \begin{bmatrix} \check{p}_1 \\ \check{p}_2 \end{bmatrix} \quad (7)$$

$$\delta_p(\mathbf{m}) = \|\mathbf{m}\|^2 \begin{bmatrix} \check{s}_1 \\ \check{s}_2 \end{bmatrix} \quad (8)$$

where $\mathbf{m} = [x \ y]^T$; $\|\mathbf{m}\|$ is the Euclidian norm of \mathbf{m} , $\|\mathbf{m}\| = \sqrt{x^2 + y^2}$; $\check{k}_1, \check{k}_2, \dots, \check{k}_n$ are the coefficients of the radial distortion; \check{p}_1 and \check{p}_2 are coefficients of decentering distortion; and \check{s}_1 and \check{s}_2 are the coefficients of thin prism distortion.

To estimate the parameters of the lenses distortion iterative methods as the bundle adjustment, are used [10]. In many applications only a second order model is used for the radial component of the lenses model as higher orders may lead to numerical instability [11]. However, lenses with higher distortions, as the fish-eye lenses, are not properly represented by simple second order models [12]. To address this issue, we propose to split the image into smaller regions and apply the method in each region. In Figure 1 we contrast the performance for different model orders and apply lenses compensation to local regions as opposed to the entire image as in traditional methods.

4. Conclusion

We investigate the effect of the lenses model order on the reconstruction error. We also propose a local method to compensate for the lenses distortion. The experiments indicate that a model up to order 6th for the radial component is enough, as reported in the literature. Moreover we find that the local method of compensation allows a reduction of average compensation error up to 3 pixels with a standard deviation of 5 pixels, while reduces the calibration error to 0.5 mm with a standard deviation of 1 mm . The proposed local method provides a better reduction of the calibration and reconstruction errors when compared to traditional methods.

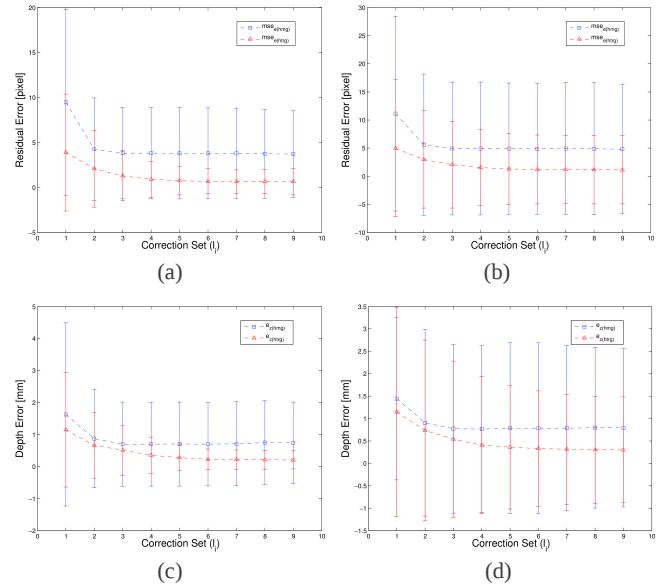


Fig. 1: Residual calibration and depth errors (for different parameters model orders 1-9) for training set (a) and (c); and for the test set (b) and (d) by considering traditional correction (blue square markers) and proposed local correction (red triangle markers). The proposed method is shown to be considerably superior than the traditional method.

References

- [1] W. Kim, A. Ansar, R. Steele, and R. Steinke, "Performance analysis and validation of a stereo vision system," in *IEEE International Conference on Systems, Man and Cybernetics*, 2005.
- [2] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.
- [3] M. PERROLLAZ, A. SPALANZANI, and D. AUBERT, "Probabilistic representation of the uncertainty of stereo-vision and application to obstacle detection," in *IEEE Intelligent Vehicles Symposium*, 2010.
- [4] F. A. Moreno, J.-L. Blanco, and J. Gonzalez, "A probabilistic observation model for stereo vision systems: Application to particle filter-based mapping and localization," in *IbPRIA*, 2007.
- [5] D. F. Llorca, M. A. Sotelo, I. Parra, M. Ocaña, and L. M. Bergasa, "Error analysis in a stereo vision-based pedestrian detection sensor for collision avoidance applications," *Sensors*, vol. 10, no. 4, pp. 3741–3758, April 2010.
- [6] T. Pinto, C. Kohler, and A. Albertazzi, "Regular mesh measurement of large free form surfaces using stereo vision and fringe projection," *Optics and Lasers in Engineering*, vol. 50, pp. 910 – 916, July 2012.
- [7] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.
- [8] D. G. Aguilera, J. G. Lahoz, and P. R. González, "An automatic approach for radial lens distortion correction from a single image," *IEEE Sensors Journal*, vol. 11, no. 4, pp. 956–965, April 2011.
- [9] D. C. Brown, "Close-range camera calibration," *PHOTOGRAMMETRIC ENGINEERING*, vol. 37, no. 8, pp. 855–866, 1971.
- [10] C. Ricolfè Viala and A. Sánchez Salmerón, "Robust metric calibration of non-linear camera lens distortion," *Pattern Recognition*, vol. 43, no. 4, pp. 1688–1699, April 2010.
- [11] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *The Proceedings of the 7th IEEE International Conference on Computer Vision*, vol. 1, 1999, pp. 666–673.
- [12] J. Wang, F. Shi, J. Zhang, and Y. Liu, "A new calibration model of camera lens distortion," *Pattern Recognition*, vol. 41, no. 2, pp. 607–615, February 2008.

Depth-Assisted Error Concealment Algorithm for 3D Video

Pin-Cheng Huang^a, Gwo-Long Li^b, Mei-Juan Chen^a, and Kuang-Han Tai^a

Dept. of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan ^a
 E-mail: m9923026@ems.ndhu.edu.tw, cmj@mail.ndhu.edu.tw, 810123002@ems.ndhu.edu.tw

Industrial Technology Research Institute, Hsinchu, Taiwan ^b

E-mail: glli@itri.org.tw

Abstract—This paper proposes an effective error concealment algorithm of whole frame loss on color information for three-dimensional(3D) video coding and transmission. The main concept of the proposed algorithm tries to extrapolate the motion vectors for concealing error block by jointly considering available motion vector and depth information spatially and temporally. Simulation results demonstrate that our proposed algorithm can achieve PSNR up to 0.57dB as well as subjective quality improvement compared to previous work.

I. INTRODUCTION

With the development of 3D video coding and communication technologies, transmitting video bit-stream through network becomes more popular. However, the video transmission may suffer from the packet loss and interference problem. Therefore, many error concealment algorithms have been proposed [1-5] to deal with this problem.

In this paper, we propose an error concealment algorithm of frame loss for 3D video coding with additionally considering depth information to select the suitable motion vector for error concealment based on [5]. By the modification of proposed algorithm, both of objective and subjective quality can be improved noticeably when compared to previous work[5].

II. PROPOSED ALGORITHM

Fig. 1 shows the flowchart of proposed motion vector selection algorithm. First, a motion vector set called MVE for extrapolated motion vectors from reference frame is constructed as shown in Fig.2 and Fig.3. However, for the intra coded blocks without motion vectors in the previously reference frame, our proposed algorithm will decide suitable motion vectors for all intra coded blocks in previously reference frame. Afterwards, the depth difference between the pixel of current 4x4 error block and the pixel pointed by motion vectors within MVE is checked to find the motion vectors with minimum depth difference. If only one motion vector satisfies the condition that its depth difference is minimum, the motion vector selection operation will be skipped and this motion vector will be assigned to $MV_{selected}$ and will be used in the following error concealment operation. However, if more than two motion vectors satisfy the condition that their depth differences are minima, the motion vector will be selected by the following equation in which both of absolute value of motion vector and the covered area are considered together.

$$MV_{selected} = \underset{MV_k \in MV_{Dmin}}{\operatorname{argmin}} \frac{|MV_k|}{\operatorname{Area}(MV_k)}, k = 1, 2, \dots, N \quad (1)$$

where N is the number of motion vectors within MV_{Dmin} which satisfies the condition that their depth differences are minima; $|MV_k|$ indicates the absolute value of motion vector, and $\operatorname{Area}(MV_k)$ means the number of pixels covered by motion vector MV_k within a 4x4 block. Once $MV_{selected}$ has been decided successfully, it will be determined to check whether its corresponding depth difference is smaller than TH or not. If the corresponding depth difference of $MV_{selected}$ is smaller than TH, the motion vector of $MV_{selected}$ will be used to conceal the current error 4x4 block. Otherwise, the depth difference surrounded by current error pixel with ± 16 search area will be searched to check whether the valid neighboring pixel can be found. Here, the valid neighboring pixels are defined by that their depth differences are smaller than 20. If none of any valid neighboring pixel can be found, the co-located motion vector of current error block in reference frame will be used to conceal the current error block directly. Otherwise, the valid neighboring pixels will be gathered to calculate the pixels to conceal current error pixel by interpolation approach.

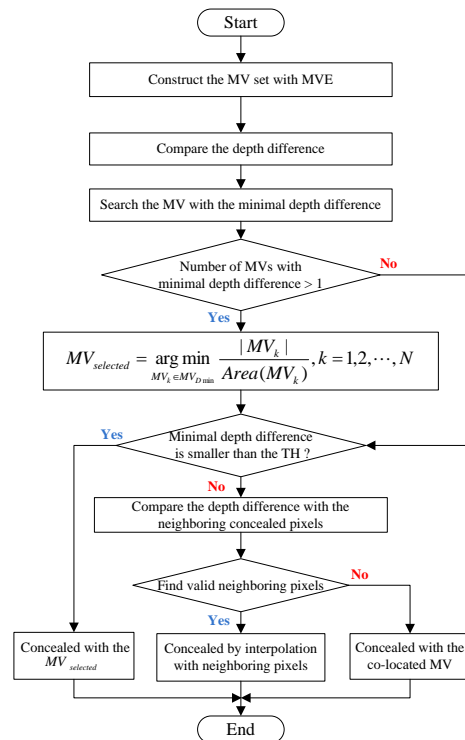


Fig. 1 Flowchart of proposed motion vector selection algorithm

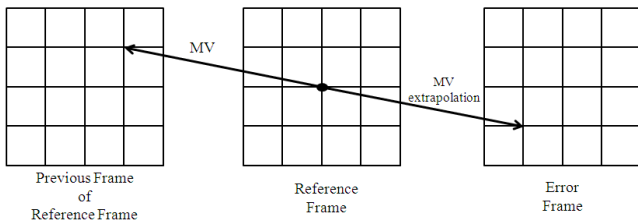


Fig. 2 Illustration of motion vector extrapolation(MVE)

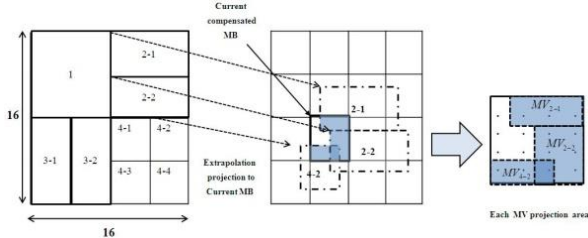


Fig. 3 Illustration of MVE construction for variable block sizes

III. SIMULATION RESULTS

This section shows simulation results to demonstrate the efficiency of our proposed error concealment algorithm. In our simulation, the reference software of JMVC 8.5 and eight test sequences are used to derive the simulation results. Furthermore, the QP is set to 22 and packet loss rate (PLR) of 5%, 10%, and 15% are tested. In addition, we assume that only the P frame packaged by single network package will be lost during the transmission process. TABLE I-II show the simulation results of our proposed algorithm. For the averaged PSNR with considering the error frames, our proposed algorithm can aim at up to 0.57 dB PSNR improvement. On average, our proposed algorithm can receive up to 0.51 dB PSNR improvement for all frames when compared to [5]. Fig.4 shows the objective comparison for Mobile sequence. We can find that our proposed algorithm can significantly improve the subjective quality.

TABLE I Average PSNR (dB) comparison between [5] and the proposed algorithm on error frames

Sequences	Mobile	Kendo	Balloons	Newspaper	LovebirdI	BookArrival	GT_Fly	Undo_dancers	Average	
Error free	44.24	45.36	44.41	42.30	42.31	41.22	42.92	41.46	43.03	
5%	[5]	35.75	29.56	33.30	35.22	38.61	32.87	34.38	27.87	33.45
	Proposed	36.32	29.73	33.53	35.39	38.63	33.00	34.62	28.29	33.69
	Δ PSNR	0.57	0.17	0.23	0.17	0.02	0.13	0.24	0.42	0.24
10%	[5]	35.03	28.53	32.90	34.32	38.32	31.19	33.01	26.44	32.47
	Proposed	35.31	28.69	33.04	34.62	38.40	31.39	33.29	26.87	32.70
	Δ PSNR	0.28	0.16	0.14	0.30	0.08	0.20	0.28	0.43	0.23
15%	[5]	34.55	27.42	32.77	34.18	37.92	30.20	31.63	26.50	31.90
	Proposed	34.88	27.65	32.95	34.41	38.01	30.29	31.85	27.06	32.14
	Δ PSNR	0.33	0.23	0.18	0.23	0.09	0.09	0.22	0.56	0.24

TABLE II Average PSNR (dB) comparison between [5] and the proposed algorithm on all frames

Sequences	Mobile	Kendo	Balloons	Newspaper	LovebirdI	BookArrival	GT_Fly	Undo_dancers	Average	
Error free	44.24	45.36	44.41	42.30	42.31	41.22	42.92	41.46	43.03	
5%	[5]	40.77	38.93	40.34	39.30	41.18	38.64	35.76	35.97	38.86
	Proposed	40.88	39.08	40.42	39.33	41.19	38.80	35.92	36.17	38.97
	Δ PSNR	0.11	0.15	0.08	0.03	0.01	0.16	0.16	0.20	0.11
10%	[5]	37.70	34.09	36.63	36.57	39.85	36.35	37.13	32.04	36.30
	Proposed	37.90	34.18	36.79	36.85	39.93	36.50	37.34	32.35	36.48
	Δ PSNR	0.20	0.09	0.16	0.28	0.08	0.15	0.21	0.31	0.18
15%	[5]	37.50	32.45	36.62	36.68	39.39	32.89	34.42	31.95	35.24
	Proposed	37.75	32.67	36.75	36.88	39.52	33.05	34.56	32.46	35.46
	Δ PSNR	0.25	0.22	0.13	0.20	0.13	0.16	0.14	0.51	0.22

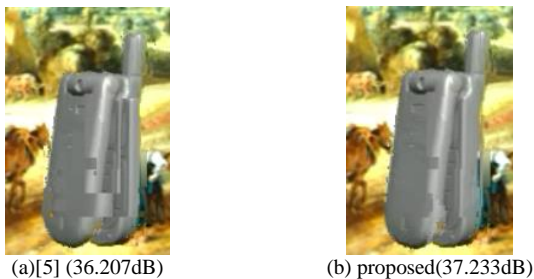


Fig. 4 Subjective quality comparison(The 123th frame of the Mobile sequence)

demonstrate that our proposed algorithm can achieve better error concealment results both in subjective and objective qualities when compared to previous work.

Reference

- [1] Bo Yan, "A novel H.264 based motion vector recovery method for 3D video transmission," *IEEE Transactions on Consumer Electronics*, vol. 53, pp. 1546-1552, November 2007.
- [2] Yunqiang Liu, Jin Wang, and Huanhuan Zhang, "Depth image-based temporal error concealment for 3-D video transmission," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 600-604, April 2010.
- [3] Tae-Young Chung, Sanghoon Sull, and Chang-Su Kim, "Frame loss concealment for stereoscopic video based on inter-view similarity of motion and intensity difference," *IEEE International Conference on Image Processing (ICIP)*, pp. 441-444, 2010.
- [4] Bo Yan and Gharavi Hamid, "A hybrid frame concealment algorithm for H.264/AVC," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 98-107, January 2010.
- [5] Bo Yan and Jie Zhou, "Efficient frame concealment for depth image-based 3-D video transmission," *IEEE Transactions on Multimedia*, vol. 14, no.3, pp. 941-945, June 2012.

IV. CONCLUSION

In this paper, we propose an error concealment algorithm for 3D video coding. By considering both of motion vector and depth information, the proposed error concealment algorithm could be able to conceal the error pixels as accurate as possible. Simulation results

CONTINUOUS GESTURE RECOGNITION USING HIDDEN MARKOV MODELS

Joceli Mayer and Vinicius Breda

Digital Signal Processing Laboratory
Federal University of Santa Catarina - Florianopolis - SC - Brazil
joceli.mayer@lpds.ufsc.br

ABSTRACT

This work presents an algorithm for recognizing gestures in videos where the actions are executed continuously without pause between them and can be performed with one or both hands. We employ Hidden Markov Models (HMM) for modeling gestures as this technique has been applied successfully in speech and character recognition. We investigate the performance for a set of 26 visual descriptors extracted from the hands after a region segmentation based on normalized quadrants. Recognition are performed by adapting the Hidden Markov Models Toolkit (HTK) and achieve a recognition rate of 91.28% for a set of 21 phrases each composed of 4 gestures from a dictionary of 15 gestures from the Brazilian sign language (LIBRAS).

Index Terms— Gesture recognition, statistical and pattern recognition, hidden Markov models.

1. INTRODUCTION

"This paper is being submitted as a poster". Hand gestures provide a universal mean of communication among people even when they do not speak the same language. A system capable of recognizing gestures can be used in many applications, such as a more "natural" interface between man and machine, for recognition of sign languages, for intrusion or dangerous/hazardous activity detection by analyzing suspicious gestures, etc. This work presents an algorithm for recognizing gestures in videos, where the actions are executed continuously without pause between them and can be performed with one or both hands. We employ Hidden Markov Models (HMM) [4] for modeling gestures, as this technique has been applied successfully in speech and character recognition. Recognition are performed by adapting the Hidden Markov Models Toolkit (HTK) which is a tool developed with a focus on speech recognition using HMMs. For extraction of the features [1, 2] required for recognition with HMMs, we propose an algorithm for the detection and tracking of the hands and face in the videos and also for the extraction of features.

2. THE PROPOSED APPROACH

The algorithm initially segments the yellow gloves (hands) and the face based on the skin color and the gloves color, both represented in the YCbCr color model as illustrated in the Figs. 1, 2 and 3. A set of descriptors (hand position, angle, direction, relative position to the other hand, derivatives of position and angle, compacticity, statistical metrics, Hu moments [3], Fourier descriptors, etc) are extracted from the frames, as in Fig. 5, which are based on their relative position to the defined and detected quadrants from the body and the face as in Fig. 4. These descriptors are trained for sub-gestures, as in Fig. 6, using the Baum-Welch algorithm (provided by the HTK tool) to define the transition matrix and emission probabilities of a set of hidden Markov models as in Fig. 7. These trained hidden Markov models are used in the Viterbi algorithm also provided by the HTK toolkit to perform the detection from a continuous video sequence.

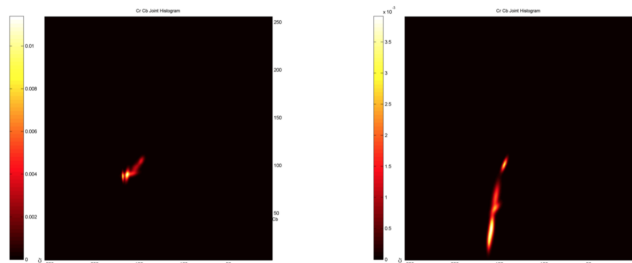


Fig. 1: Face and hands segmentation is based on the color distributions for the face (on the left) and for gloves (on the right).

3. EXPERIMENTS

For training the system, 81 phrases composed of 4 gestures (form a set of 15 gestures from the Brazilian sign language - LIBRAS) were performed in a controlled manner using yellow gloves and a setting with white background. A set of 26 descriptors and the structure of HMMs (number of states and number of Gaussian functions necessary to model the prob-

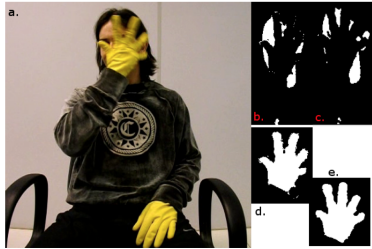


Fig. 2: Face and hands segmentation with occlusion.

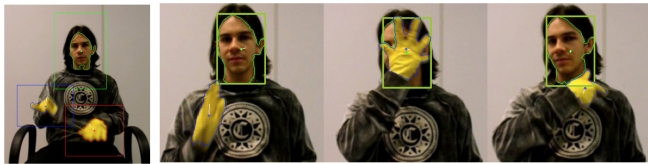


Fig. 3: Face and hands segmentation with located center in a boundary box and occlusion.

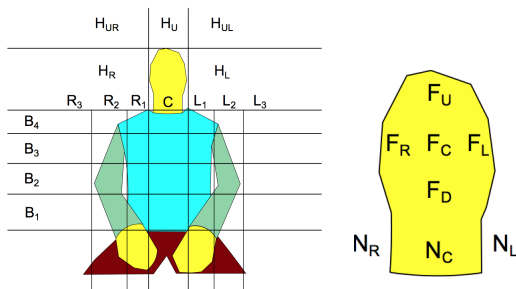


Fig. 4: Defined quadrants and defined regions on face.

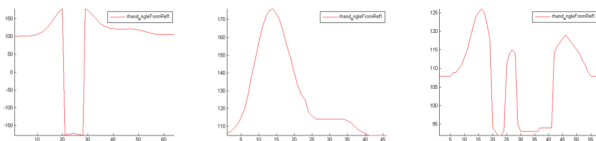


Fig. 5: Evolution in 3 frames of the descriptor based on the angle of the hand to the reference.

ability of emission of symbols from the states) are proposed by optimizing the number of descriptors and number of Gaussian function based on larger set of features. This set of descriptors and the defined structure of HMMs are evaluated by observing the resulting performance of the gesture recognition system for a set of testing videos. The accuracy achieved was about 91.28% of correct classification of 21 phrases composed of 4 gestures from a dictionary of 15 gestures from the Brazilian sign language (LIBRAS). We employed cross-validation tests and found that the rate of correct recognition of gestures ranged from 86% to 97% for different sets.

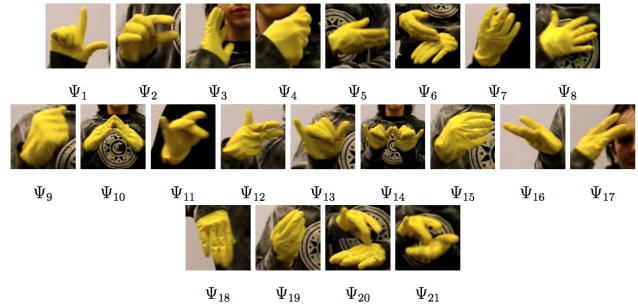


Fig. 6: Examples of sub-gestures that form complete gestures.

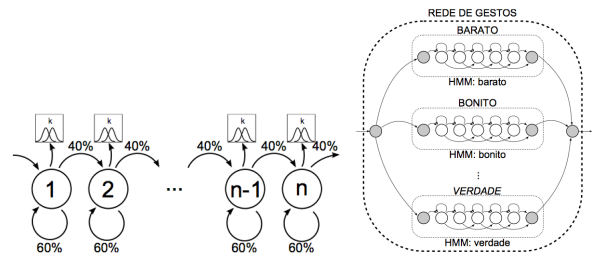


Fig. 7: Hidden Markov model for a sub-gesture and the network of Hidden Markov models to build a set of gestures.

4. CONCLUSIONS

We propose an algorithm for detection of gestures. The algorithm consists of labelling, segmentation, extraction of features based on defined body quadrants. The hidden Markov models provide an efficient approach to model the dynamics of the descriptors (features) in the video sequences. We achieve an accurate classification for a dictionary of 15 gestures from the Brazilian sign language (LIBRAS).

5. REFERENCES

- [1] J. Flusser and T. Suk. Pattern recognition by affine moment invariants. *Pattern recognition*, 26(1):167-174, 1993.
- [2] L. Gupta and M. D. Srinath. Contour sequence moments for the classification of closed planar shapes. *Pattern Recognition*, 20(3):267-272, 1987.
- [3] M.-K. Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179-187, 1962.
- [4] A.D Wilson and A.F. Bobick. Parametric Hidden Markov Models for Gesture Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9), 1999.

Local Block-Difference Pattern for Use in Gait-based Gender Classification

Yen-Chi Wang, Chiao-Wen Kao, Ying-Nong Chen and Kuo-Chin Fan

Institute of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan

Abstract - In this paper, a novel local texture descriptor termed as Local Block-Difference Pattern (LBDP) is proposed. In conventional LBP, sensitive to intensity change problem will drastically affect the performance due to its simple pixel value comparison mechanism. Different from LBP, the proposed LBDP describes the local textures from a pixel to a block for decreasing the impacts resulting from intensity change. Based on the proposed LBDP, the tolerance to the intensity change is exaggerated because of the expanding of encoding range. The effectiveness of the proposed LBDP is practically demonstrated in the application of gait-based gender classification. In the experiments, CASIA dataset B is adopted and the experimental results demonstrate that the proposed LBDP outperforms the other LBP-based descriptors.

Keywords: Local texture, LBP, LBDP, Gender classification

1 Introduction

Recently, gender classification attracts a lot of researchers due to its tremendous applications including surveillance, face recognition, video indexing, and marketing survey system. One typical application is intelligent surveillance system. In intelligent surveillance system, gender is used as a biometric feature to prevent the intrusion of inappropriate gender in inappropriate places (such as lady room) which enhances the practicality and add-on value of traditional surveillance systems. Therefore, abundant of literatures were presented [1]-[4] for gender classification. However, most of them used frontal face images or voice for classification. Even the frontal face images can be obtained, the classification performance is easily affected by pose, illumination, and expressions. From the above description, the most critical issue in gender classification is feature extraction. Raised by this motivation, we propose a novel texture descriptor to extract useful information for gender classification. The proposed scheme is based on LBP method [5], termed Local Block-Difference Pattern (LBDP) is proposed. The aim of the proposed LBDP is to alleviate the intensity change and improve the discriminative ability. To this end, we encode the textures into binary codes via blocks instead of pixels. To evaluate the effectiveness of the proposed scheme, LBDP is applied on gait-based gender classification. CASIA dataset B is a benchmark which used for gait-based gender classification. Gait Energy Images (GEI) scheme is firstly applied on silhouettes images. Then, the

proposed LBDP is applied on GEI to extract discriminant texture features. Finally, support vector machine is adopted to train the gender classifier. The main contribution of this study is that the proposed LBDP is a novel texture descriptor and successfully applied in gait-based gender classification. The main reason is that LBDP can alleviate the sensitive to intensity change problem of LBP and could obtain more powerful discriminative pattern. Moreover, LBDP is also computational effective. The rest of this paper is organized as follows: Algorithm local block-difference pattern is presented to obtain the block-based texture in section 2. In section 3, experiments are conducted to show the effectiveness of the proposed method. Finally, conclusions are given in section 4.

2 Local Block-Difference Pattern

In the proposed LBDP, we firstly take each block as a vector. Then, a given center vector is compared to its p neighboring vectors with radius R . To remedy this problem and enhance discriminative ability, we estimate the similarity of center block and its neighbor block replace LBP encode schema. LBDP take center block as a vector so that utilized the element of the vector to vote the similarity between each blocks. Based on voting schema, the effect of intensity change can be decreased. As to discriminative ability of LBDP, like MB-LBP, the encoding strategy is to describe the structure of local blocks instead of pixel. The basic idea of proposed LBDP is that if the element of neighboring vector is bigger than correspond element of center vector, the element is recorded as 1, 0 otherwise. Finally, the binary code of the neighboring vector is determined by a voting mechanism, if the number of "1" value is bigger a threshold, the binary code of this neighboring vector is set as 1. Otherwise, it is set as 0. The LBDP can be calculated by Equations (1) and (2).

$$LBDP_{p,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(B_p - B_c) \times 2^p \quad (1)$$

$$s(x) = \begin{cases} 1 & \left\| \sum_{i=1}^{blocksize} (B_{p_i} - B_{c_i}) > 0 \right\| \geq threshold \\ 0 & otherwise \end{cases} \quad (2)$$

Where B_c is a given center block, and B_p is a neighboring block with radius R from B_c . The given center block can be encoded into a P -bit binary code via concatenating the binary

codes of its neighboring blocks. In our experiments, P is set as 8 to generate a 2^8 binary feature. LBP focuses on the center point and its neighboring points. It means that LBP only concerns the ration between each point in the center block. Since the encoding strategy of LBP is pixel-based, noisy signal would tend to be embedded. Due to this reason, we extent the ration from point-based to block-based and combine with voting scheme. Hence the tolerance ability of intensity change can be improved. Another issue is how to determine the number of neighbors. The number of neighbors in LBDP is fixed to 8. The most difference between LBDP and other local texture descriptors is that the block size and overlapping area.

3 Experimental results

In this section, gait-based gender classification is conducted to show the effectiveness of the proposed LBDP. We compare our method with LBP, LDiP, LDeP and MB-LBP these four type of local texture descriptors. The benchmark database, CASIA dataset B is used for evaluating the classification performance. In the experiments, 31 males and 31 females are chosen to conduct gait-based gender classification. The cross-validate strategy is used in this study, one male and one female are used as testing samples and the others are used as training samples. Before the comparison of classification performance, the procedures of gait-based gender classification in the experiment is described. First, foreground objects from training images are detected by using a simple background subtraction method. Then, auto-threshold method is utilized to extract the binary images of foreground objects. To remedy scale problem, the foreground object is normalized to a fix size of 160 by 100. Next, gait energy images (GEI), a very powerful and useful gait descriptor with not only low computation cost but also preserving both static and dynamic information is applied to describe human gait. Finally, the proposed LBDP is applied to extract block-based texture from GEI and feed Support Vector Machine (SVM) classifier for gender classification. The performance comparison of different local texture descriptors is shown as table1. We not only compare the four local texture descriptors but also compare other method from the state-of-art. At last we can find out that the proposed method achieves the better performance than almost other methods. In the table, our proposed method is worse than Hu. But according Hu's methods, they need more computation cost. Our proposed method is more efficiently and lowly computation cost.

4 Conclusions

In this paper, a new local texture descriptor is proposed. It can enhance useful information and discriminative ability. Local texture descriptors, LBP, LDiP, LdeP and MB-LBP are used to estimate the performance of LBDP. The experimental results have shown that LBDP provided the more

discriminative power than other local texture descriptors. LBDP can use for gait-based gender classification.

Table 1: Comparisons of different algorithms

Local texture descriptors	Recognition rate
LBP	95.96
LDiP	95.69
LDeP	95.96
MB-LBP	96.51
Li et al.[6]	93.28
Yu et al.[7]	95.97
Hu et al.[8]	96.77
Hu et al.[9]	98.39
The proposed LBDP	97.31

5 References

- [1] J.Bekios-Calfa, J.M.Buenaposada and L.Baumela, "Revisiting Linear Discriminant Techniques in Gender Recognition," *IEEE Transactions on. Pattern Analysis and Machine Intelligence*, Vol.33, No.4, pp. 858-864, 2011.
- [2] B. Moghaddam and M.-H. Yang, "Learning Gender with Support Faces," *IEEE Transactions on. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, pp. 707-711, May 2002.
- [3] S. Baluja and H.A. Rowley, "Boosting Sex Identification Performance," *International Journal of Computer Vision*, Vol. 71, No. 1, pp. 111-119, Jan 2007.
- [4] E. Ma "kinen and R. Raisamo, "Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces," *IEEE Transactions on. Pattern Analysis and Machine Intelligence*, Vol. 30, No. 3, pp. 541-547, Mar 2008.
- [5] T. Ojala, M. Pietikäinen, and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Feature Distributions", *Pattern Recognition*, Vol. 29, pp. 51-59, 1996.
- [6] X. Li, S. J. Maybank, S. Yan, D. Tao, and D. Xu, "Gait components and their application to gender recognition," *IEEE Transactions on System, Man and Cybernetics – part C*, Vol. 38, No. 2, pp. 145-155, 2008.
- [7] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Transactions on Image Processing*, Vol. 18, No. 8, pp. 1905-1910, 2009.
- [8] M. Hu, Y. Wang, Z. Zhang, and Y. Wang, "Combining spatial and temporal information for gait based gender classification," in *Proceedings of International Conference on Pattern Recognition*, pp. 3679–3682, Aug 2010.
- [9] M. Hu, Y. Wang, Z. Zhang and D. Zhang, "Gait-Based Gender Classification Using Conditional Random Field," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.41, No.5, pp. 1429-1439, 2011.

Comparative Analysis of the Optimization Techniques of Image Filters Used to Improve Optical Character Recognition on Text Images

Iulia Știrb

Computer and Software Engineering Department, "Politehnica" University of Timișoara, România

Abstract - Image filtering is changing the appearance of the image by altering the shades and colors of the pixels. Increasing the contrast as well as adding a variety of texture, tones and special effects to images are some of the results of applying image filters for the purpose of improving the success rate of OCR (Optical Character Recognition) performed on text image. Still the main purpose of filtering is reducing the noise around characters in the image. Within Silicon and Software System Limited (S3Group) Company from Dublin, Ireland, the image filters described in this paper are used to improve a subsequent OCR which is part of the process of testing the menu options displayed on the output video of the set-top box under test. This paper focuses on the optimization techniques and on the analysis of which better suits to be applied to each of the image filters in order to make them perform faster. The optimization that produces the best improvement in terms of execution time for all filters above is done using a byte array to store the components i.e. Blue, Green, Red and Alpha of each pixel. Two other optimizations are described, one using C# predefined `ColorMatrix` class for improving Contrast filter and the other by computing just once the filter weight (i.e. sum of all template elements) for all pixels in the interior of the image in the case of Sharpen and Blur filters, which both use a template to compute the new value of the pixels.

Keywords: image, filter, optimization, speedup, contrast, OCR

1 Introduction

The OCR of videos, involving first an automatic extraction of many frames (captures) in a second, as in [1], is proved to be very useful in digital television domain. Lately, in digital television, the need for a high number of OCR on text images in a short time interval is increasingly common. Filtering the images before passing them to OCR is becoming more and more frequent because of the need to remove the noise. Since the OCR speed must increase, so the filtering speed must do.

The optimization techniques in terms of execution time of Contrast, Sharpen, Blur, Invert, Color and Highlight filters are presented the first and second sections of this paper and the techniques will be compared in the third section. Snippets of initial and optimized (using byte buffer) implementation of image

filters are listed in the fourth section. Conclusions section states the proper optimization technique to be used for each filter.

What this paper brings new is a case study of which optimization technique suits well to each filter separately. I have carried out many tries of applying the optimization techniques on all filters. For instance, Contrast filter's most optimized version is the one in which the implementation is done using `C# ColorMatrix` class.

2 Optimization Techniques

2.1 Usage of a byte buffer

The byte buffer stores byte representation of the image, namely Blue, Green, Red and Alpha components in this order for each pixel as in Fig. 1. An additional cloned buffer is required for Sharpen and Blur filters to avoid altering the original pixels, since the new value of the current pixel depends on its neighboring pixels.

First step in the optimization algorithm is to obtain an object of type `BitmapData` (more details can be found in [2]) using `C# LockBits` method of `Bitmap` class. In next step `Scan0` property is accessed on the obtained object in order to get the address of the first pixel in the image. Once we have this address, all the image data from it is copied to the byte buffer using the overloaded `Copy` method of `C# Marshal` class. The third step is different from a filter to another and represents the calculations done on the byte buffer in order to filter the image. Finally, the byte buffer will contain the byte representation of the filtered image. By using the buffer, we avoid calling `SetPixel` method on a `Bitmap` object each time the pixels needs to be set to the filtered value, which would be time consuming since the `Bitmap` object is accessed as many times as the number of pixels in the image. The next step is to copy the filtered buffer back to the address of the first pixel of the image using the overloaded `Copy` method. The last step is to unlock the `Bitmap` object, using `UnlockBits` method.

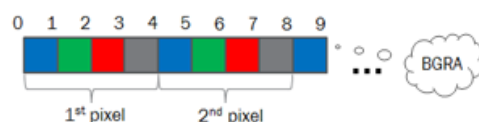


Figure 1. Image representation storage in the byte buffer

2.2 Usage of ColorMatrix class

Figure 2 present an example of how this technique performs for one pixel with e.g. Red component equal to 51 and scale down to 0.2. As in [3], the input vector represents the original pixel with its components Red, Green, Blue and Alpha in this order. The values of the components vary in a range from 0 (= 0 / 255) to 1 (= 255 / 255). The last element of the input vector is used for additional computations, if needed. After the calculation of the output array, the values are scaled again to the fit into the interval [0;255].

The matrix content is specific to each filter and exemplified in Fig. 3 for Contrast filter. *diff* is computed once for all pixels as in (1) and *scale* is the desired degree of Contrast (usually varies from 1 to 6). *diff* usage saves a lot of calculations per pixel, that should have normally been made.

$$diff = -0.5 * (scale - 1.0) - 0.5 / 255.0 \quad (1)$$

2.3 Comparison Between the Optimization Techniques

Since using `ColorMatrix` implies having to perform the operation on the `Bitmap` object that encapsulates the image, using also a byte buffer would make no sense, so the two cannot be both applied to image filters. Details can be found in Table I.

TABLE I. Comparison between the optimization techniques improvements in terms of execution time of the image filters

Curr. No.	Image Filter	Optimization technique used			Speedup
		Byte Buffer	Color Matrix	Filter Weight	
1	Contrast	-	✓	-	8.7 times
2	Sharpen	✓	-	✓	4.7 times
3	Blur	✓	-	✓	4.7 times
4	Invert	✓	-	-	53 times
5	Color	✓	-	-	42 times
6	Highlight	✓	-	-	16 times

2.4 Initial Implementation of Filters and the One Optimized Using the Buffer

Code 1. Image filters implementation before being optimized (C#)

```
Bitmap clonedBitmap = new Bitmap(originalImage.Width,
                                originalImage.Height);
UnsafeBitmap bitmap = new UnsafeBitmap(originalImage);
bitmap.LockBitmap();
...
// for each pixel with coordinates (x,y) set the filtered value
clonedBitmap.SetPixel(x,y,Color.FromArgb(r,g,b));
...
bitmap.UnlockBitmap();
return clonedBitmap;
```

Code 2. Image filters optimized implementation using the byte buffer technique (C#)

```
FilteredBitmap filteredBitmap = new
    FilteredBitmap(originalImage);
byte[] buffer = filteredBitmap.LockBitmap();
...
(0.2 0.7 0.5 1.0 1.0) ×  $\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0.5 & 0.1 & 0.2 & 0 & 1 \end{pmatrix}$  = (0.9 0.8 0.7 1.0 1.0)
```

Figure 2. `ColorMatrix` example of multiplication for one pixel

scale	0	0	0	0
0	scale	0	0	0
0	0	scale	0	0
0	0	0	scale	0
diff	diff	diff	diff	1

Figure 3. Color matrix for Contrast image filter

```
// here is the code which processes the buffer elements obtaining
// the filtered values; this part is different from a filter to another
...
// set the new value of all pixels, by accessing the image only once
filteredBitmap.SetPixels(buffer);
filteredBitmap.UnlockBitmap();
return filteredBitmap.Bitmap;
```

In Code 1, a reason for the high execution time when using `SetPixel` C# method is the assumption that the image representation is entirely loaded in memory each time a pixel is set with the new color (r,g,b).

The code which is optimized using byte buffer technique is presented in Code 2. `FilteredBitmap` is the class I have implemented which hides all the details. By using this technique, the image will be accessed only once for setting all pixels with the filtered values, considerably improving the execution time.

3 Conclusions

Filters can be combined in many ways to improve a subsequent OCR on text images. If there is a lot of noise in the image, a Blur filter is the proper solution, afterwards followed by Color filter. As in [4], this success rate of OCR also depends on whether the characters are broken or they touch each other.

Implementing the proper optimization technique for each filter depends on the particularities of the filters:

- The usage of a byte buffer is the technique with the largest applicability over image filters such as Invert, Color, or even Sharpen and Blur. The optimization is even greater as the filter perform a smaller number of computations, which is specific to the majority of image filters
- `ColorMatrix` is the most desired solution when the filter performs a lots of calculations such as Contrast filter and the calculations can be refactored and thus, reduced to a smaller and usually fixed number of computation involved when using `ColorMatrix`; this optimization cannot be applied for filters for which use a template to compute the filtered pixels

4 References

[1] T. Sato, T. Kanade, E.K. Hughes, M.A. Smith, *Video OCR for digital news archive*, Carnegie Mellon University, Pittsburg, United State, 1998

[2] "Bitmap Class", Microsoft, 2014, [Online: [http://msdn.microsoft.com/en-us/library/system.drawing.bitmap\(v=vs.110\).aspx](http://msdn.microsoft.com/en-us/library/system.drawing.bitmap(v=vs.110).aspx)]

[3] "ColorMatrix Class", Microsoft, 2014, [Online: [http://msdn.microsoft.com/en-us/library/system.drawing.imaging.colormatrix\(v=vs.110\).aspx](http://msdn.microsoft.com/en-us/library/system.drawing.imaging.colormatrix(v=vs.110).aspx)]

[4] P. Stubberud¹, J. Kanai, V. Kalluri, "Adaptive image restoration of text images that contain touching or broken characters", ¹Nevada University, Las Vegas, United States of America, 1995

INCREASING THE CAPACITY OF COLOR PRINT CODES FOR ROBUST COMMUNICATION OVER INKJET PRINT-SCAN CHANNELS

Joceli Mayer

Digital Signal Processing Laboratory
Federal University of Santa Catarina - Florianopolis - SC - Brazil
joceli.mayer@lpds.ufsc.br

ABSTRACT

This investigation on robust and high capacity print codes aims to increase information payload in a given printed page area while providing robustness to channel errors including distortions originated by the inkjet printing and scanning processes. The approach includes statistical print-and-scan channel characterization, designing of robust segmentation using visual cues, unsupervised Bayesian color classification with expectation-maximization algorithm for parameters estimation of a mixture of Gaussians model and design of error correction codes. Results illustrate the performance evaluated under real channel and distortions conditions. High payload of 4592 bytes per squared inch is achieved with a robustness of 92% to distortions due to the print-and-scan channel. Adding high-density information to printed materials enables interesting hardcopy document applications involving security, authentication, physical-electronic round tripping, item-level tagging, and consumer/product interaction.

Index Terms— Statistical pattern classification, hardcopy watermarking, document authentication.

1. INTRODUCTION

”This paper is being submitted as a poster”. Most prior work on printed codes is restricted to 1D barcodes that have limited payload. A recent investigation [1] has shown a great potential of two-dimensional printed code technologies to improve information transmission using paper. Related research on barcodes aimed to be captured by cell phones or consumer digital cameras has been proposed in the literature [2, 3] and some results can be applied to our research that involves acquisition by scanners. Recent studies such as [1, 4, 5] have investigated the use of 2D patterns to improve capacity and robustness. The literature clearly indicates that significant improvements in capacity and robustness can be achieved by exploiting color and 2D patterns. Achieving robust high-capacity information transmission over the print-and-scan channel requires addressing of a variety of degradations. These include distortions due to: i) Noise and the optical blurring in the channel; ii) Geometric disturbances; iii) Ink

fading, spills, creases and aging; iv) Document manipulation by the user. Novel error correction coding, statistical classification and robust segmentation techniques need to be investigated to address these distortions. Another challenge is to develop technologies that are robust enough for general purpose use over a variety of embedding (printing) and detection (scanning) devices. By interacting with the paper using 2D color codes, new and interesting applications will be possible. Some examples include the restoration of aged printed photos with the help of 2D color codes printed on the back of the photo [6], content authentication using barcodes confined to small areas, security printing with smart labels [7, 8], and deterrents for branded product counterfeiting. We report recent results to achieve higher capacity and robustness based on the approach of our previous paper [9]. In this paper we employed 32 colors, as opposed of only 4 color in [9], for the print codes and improved segmentation and detection techniques to achieve a high payload of 4592 bytes per inch squared (in^2), as opposed to 2500 bytes/ in^2 in [9], while providing a very high robustness of about 98% percent.

2. THE PROPOSED APPROACH

The communication channel includes the inkjet print-and-scan channel (PS) and external distortions. This channel includes linear and non-linear distortions: electronic, ink and paper noise; optical and motion blurring in the scanner device, ink print spreading and geometrical disturbances. The inkjet print process spreads the ink dots and the color varies considerably more than laserjet print channels.

Similarly to [9], which employed only 4 colors, we employ a Bayesian approach [10] to classify each printed and scanned block pattern in one of the 32 colors or classes. A feature vector $\mathbf{y}_j = f(\mathbf{x} \in B_j)$ is extracted from the acquired 3-dimensional samples within the j -th block pattern B_j . \mathbf{y}_j is a D -dimensional vector which is a function of the samples within B_j . Consider the class conditional probability density function (pdf), $p(\mathbf{y}|\omega_i)$, of the feature vectors \mathbf{y} belonging to class ω_i , where $i = 1, \dots, 32$. Define $P(\omega_i)$, $i = 1, \dots, 32$ as the prior probability. Thus, the posterior class probabilities

are given by [10]

$$P(\omega_i|\mathbf{y}) = \frac{p(\mathbf{y}|\omega_i)P(\omega_i)}{\sum_{j=1}^4 p(\mathbf{y}|\omega_j)P(\omega_j)}. \quad (1)$$

The posteriors provide information about which is the most probable embedded color for a given pattern block. As in [9], we assumed a Normal distribution for each class. Thus, the class conditional probability density function for a given class ω_i is given by:

$$p(\mathbf{y}|\omega_i) = \frac{1}{(2\pi)^{3/2}|\mathbf{C}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{y}-\boldsymbol{\mu}_i)\right\} \quad (2)$$

The covariance matrices, \mathbf{C}_i , the mean vectors, $\boldsymbol{\mu}_i$ and the prior probabilities $P(\omega_i)$, $i = 1, \dots, 32$ need to be estimated given the observed print codes.

Assuming the data is distributed according to a mixture of K Gaussians,

$$p(\mathbf{y}) = \sum_{k=1}^K P(\omega_k) \frac{1}{(2\pi)^{D/2}|\mathbf{C}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_k)^T \mathbf{C}_k^{-1}(\mathbf{y}-\boldsymbol{\mu}_k)} \quad (3)$$

We employ the **Expectation-Maximization** (EM) algorithm [10] to estimate the parameters of (3). The 32 posteriors are computed using the estimated parameters. The color embedded at block B_j is classified in the class i that results in the greatest $P(\omega_i|\mathbf{y})$. In the absence of user induced degradations, the unsupervised estimation with EM and Bayesian classification provides an optimal framework to deal with print and scan channel disturbances.

The employed segmentation is described with details in [9]. Basically the proposed segmentation approach relies on morphological operations to estimate of the 4 corners of the print codes bounding box followed by creating a grid to separate individual blocks. The resulting grid is robust to channel noise, internal scribing, coffee spills and geometric distortions. We also employ auxiliary visual cues to address mechanical misalignments, as proposed in [9].

We propose to employ more than 4 colors, which are typically used and propose to choose these colors from a major set, which provides improvements over 10% in robustness. We select the 32 color from a set of 256 possible colors by estimating the colors that are more distant considering a Mahalanobis distance: the candidates colors are printed and scanned and their distances are estimated, the 32 more distant colors are selected.

3. CONCLUSIONS

In this paper we proposed an approach to select 32 colors from a set of 256 colors for the print codes and improved segmentation and detection techniques to achieve a high payload of 4592 bytes per inch squared (in^2) while providing a very

high robustness of about 98% percent. . Most works in literature employs only 4 colors and achieve a robust but inferior payload of 2500 bytes/ in^2 as in [9].

4. REFERENCES

- [1] R. Villan; S. Voloshynovskiy; O. Koval; J. Vila; E. Topak; F. Deguillaume; Y. Rytsar and T. Pun, "Text data-hiding for digital and printed documents: theoretical and practical considerations," in *Proc. of SPIE*. SPIE, 2006, pp. 15–19.
- [2] D. Parikh and G. Jancke, "Localization and segmentation of a 2d high capacity color barcode," in *IEEE Workshop on Applications of Computer Vision*, 2008, pp. 1–6.
- [3] Songwen Pei; Guobo Li and Baifeng Wu, "Codec system design for continuous color barcode symbols," in *IEEE 8th International Conference on Computer and Information Technology Workshops*. IEEE, July 2008, vol. 8:11, pp. 539 – 544.
- [4] D. Shaked; Z. Baharav; A. Levy; J. Yen and N. Saw, "Graphical indicia," in *Proc. IEEE International Conference on Image Processing*. IEEE, September 2003, vol. 1, pp. 485–488.
- [5] N. Damera-Venkata and J. Yen, "Image barcodes," in *Proc. SPIE Color Imaging VIII: Processing, Hardcopy and Applications*. SPIE, January 2003, vol. 5008, pp. 493–503.
- [6] R. Samadani and D. Mukherjee, "Photoplus: Auxiliary information for printed images based on distributed source coding," in *Visual Communications and Image Processing*, January 2008.
- [7] S. Simske; J. Aronoff; M. Sturgill, "Spectral pre-compensation of printed security deterrents," in *The conference on optical security and counterfeit deterrence*, January 2008.
- [8] S. Simske; J. Aronoff; M. Sturgill; G. Golodetz, "Security printing deterrents: A comparison of thermal inkjet, dry electrophotographic and liquid electrophotographic printing," in *J. Imaging Sci. Tech.*, October 2008, vol. 52:5, pp. 1–7.
- [9] Joceli Mayer; Jose Carlos M Bermudez; Andrei P Legg; Bartolomeu F Uchoa-Filho; Debargha Mukherjee; Amir Said; Ramin Samadani; S Simske, "Design of high capacity 3d print codes aiming for robustness to the ps channel and external distortions," in *16th IEEE International Conference on Image Processing*, 2009, pp. 105–108.
- [10] C. Bishop, "Pattern recognition and machine learning," in *Springer*, 2006.

A Bluetooth Based Mobile SW Platform Application: Smart Finder

Hong-Mok Choi, Jaejoon Kim

School of Computer and Communication Engineering, Daegu University, Gyeongsan, Korea

Abstract - With the development of smart high technology related terminals, people's life style has been gradually changed. While electronic devices become smaller and smaller, people often misplace their belongings. This can arise to seriously problems for the person in case of valuable information lost. However, most smart devices have Bluetooth capability. Bluetooth is a radio networking protocol in a short range to communicate electronic terminals wirelessly. In this paper we propose a system, smart finder, to find misplaced objects based on Bluetooth technology. We reviewed Bluetooth technology and then designed hardware and software architecture for smart finder application. It has been shown the possibility and capability for many applications.

Keywords: Bluetooth, smartphone application, RF communication, smart finder

1 Introduction

As of 2012, research shows that 63.7% of the current population owns a smart device and that the average household has 0.64 of these devices. This dramatic increase shows that smartphones can provide effective information processing for daily activities [1]. Most smartphones have Bluetooth capabilities that can provide easy wireless connections. In the current complex society cognitive abilities and memory is becoming a major interest. Many problems can arise from the inability to remember. For example if you lose your car keys, wallet, phone etc. you use time to find those objects. If you lose a USB or something that contains valuable information, that information can be used against you. We explored this project to prevent from accidents as mentioned above.

2 Related Research

2.1 Bluetooth

While there has been an increase in smartphones that have NFC (Near Field Communication), there has not been an increase in applications that use NFC. Google enabled the use of their Google Wallet app without the use of NFC [2]. Unlike NFC that is limited in use of terminals, Bluetooth is supported by various terminals and is a short range, low power, low cost wireless interface. It uses 2.4GHz, ISM (Industrial Scientific Medical) frequency band which does not require a separate

license to use. Also the 1MHz bandwidth is divided into 79 channels and uses FHSS (Frequency Hopping Spread Spectrum) to hop through the channels. The range of Bluetooth is from 10 - 100m depending on the power class and the fastest transfer speed is 1Mbps (723.2kbps) [3]. We listed some prominent features of Bluetooth below

- Low energy cost
- From Bluetooth 3.0, transfer speed of 24Mbps has been supported. Possibility of short distance wireless LAN connections using the advantages over massive data transfer.
- From Bluetooth 4.0, it has been reported that the energy usage has been reduced upto 100 times.

2.2 Lost protector

As the number of electronic devices has increased, we looked to create an App that could prevent missing items. We used RF which is wireless to ring an alarm when the distance increases beyond the set limit between the user and object.

3 Implementation Design

For this research we attached a reduced H/W to the missing object. The final goal of this research is to integrate the compact tag to Bluetooth communication and distribute commercially.

3.1 Experimental system and H/W design

The smartphone application we used in this research was developed using Eclipse development tool¹. H/W coding was done through AVR Studio 4.1 and the Bluetooth module and MCU parts were manufactured by Firmtech Corp's FB155BC² and atmega128³ respectively. Figure 1 shows a specification of BT configuration that is one of utilities provided by Firmtech Corp. for Bluetooth module setup. By assigning the role as "slave" and the smartphone as "master", we allowed the smartphone to find missing items.

As the Bluetooth module FB155BC requires 3.3V, we need a regulator that matches it. We selected a LM3940⁴ regulator that can lower 4.5V from three basic AAA batteries to 3.3V. In order to find a missing item, we used GEC-33A as

¹ <http://www.eclipse.org/jdt>

² <http://firmtech.co.kr>

³ <http://www.atmel.com>

⁴ <http://www.ti.com/lit/ds/symlink/lm3940.pdf>

"This paper is being submitted as a poster".

a buzzer for the user. We used module MAT128-100 circuit as the MCU in order to experiment using the breadboard. Figure 2 shows how the MCU works. We first initialized UART (Universal asynchronous receiver/transmitter) because the Bluetooth module and the MCU communicate using the UART [4]. After initialization it operates in an infinite loop according to the transmitted character. If the value is null, it means that no character was transmitted thus the loop runs again until '0' or '1' is returned. Depending on the value the GPIO of Atmegal28 turns the buzzer on or off.

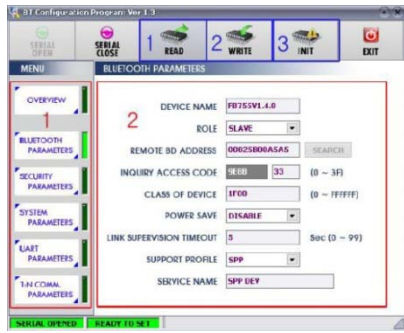


Fig. 1. Specification of BT Configuration provided by Firmtech Corp

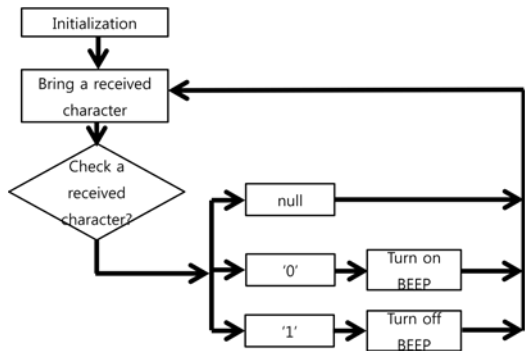
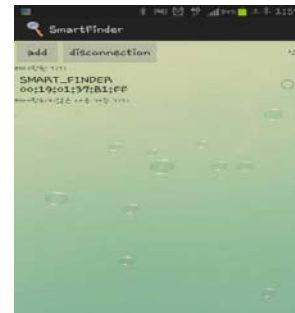


Fig. 2. Design diagram for MCU operation function

4 Discussion and Conclusion

This paper implemented a smart finder application based on Bluetooth technology. Though it was a simple design for the experiment, we think it can be useful. Also there needs to be competition between the final product size and price. If we consider that the Bluetooth range is 10m it may be difficult to find the object in a large room. Although we used Bluetooth for our experiment we advise for future works to use a different wireless communication method to increase the search range. In this experiment we relied on the user to find the object using sound, but as sound decreases over longer distances some modifications will be needed for commercial use.



(a)



(b)

Fig. 3: Initial experimental prototypes for smart finder application, (a) GUI and (b) H/W configuration

3.2 Application design

By assigning the roles, the smartphone and Bluetooth of H/W can communicate through the Bluetooth socket. By assigning the UUID as "00001101-0000-1000-8000-00805F9B34FB" it can communicate with H/W through serial communication. The GUI of the smartphone application is showed in Figure 3(a). The add button prints out the Bluetooth enable devices nearby. The list has both paired devices and unpaired devices and by clicking on a device the phone can start communicating with the selected device. If it is successful H/W receives a value of '1' and in the status is changed to "connected". If the communication is unsuccessful then a error message is shown. By communicating the H/W sounds the buzzer and the person finds their missing object, they can disconnect Bluetooth. By disconnecting it the buzzer from H/W is turned off. Figure 3(b) shows a completed initial design.

References

- [1] Lim, J M, Yoo, J Y, Jang, S J, Yoo, J M, Kim, S H, "Survey on the Internet Usage. Internet Statistics Report of 2012"; Korea Internet Statistics Information System.
- [2] H. Aziza. "NFC Technology in Mobile Phone Next-Generation Services"; 2010 Second International Workshop on Near Field Communication (NFC), pp. 21-26, 2010.
- [3] Pravin Bhagwat. "Bluetooth: Technology for Short-Range Wireless Apps"; IEEE Internet Computing, pp. 96-103, 2001.
- [4] Adam Osborne. "An Introduction to Microcomputers Volume 1: Basic Concepts". Osborne-McGraw Hill, pp. 116-126, 1980.

Facial Expressions From Raspberry Pi and Pi Camera

M. Chapron¹, F. Adon¹, J.B. Liagre¹

¹ETIS, ENSEA, UCP, CNRS

6 avenue du Ponceau, 95014 Cergy-Pontoise, France

chapron@ensea.fr

Abstract- Herein, the work proposed uses the Raspberry Pi B in order to recognize in real time the facial expressions with the use of the language C++ and openCV library. Viola Jones algorithm from openCV permits to track the face, then the rectangles around the eyes and mouth are automatically computed from the rectangle given by Viola Jones procedure. Local Binary Pattern (LBP) is computed from the sampled pixels in mouth and eye areas. Statistical parameters are estimated during the learning phase according to the different facial expressions, then they are used to classify the facial expressions during the on line stage.

Keywords: Raspberry Pi, facial expression recognition, LBP, parameter estimation.

1 Introduction

The Raspberry Pi (R.P.) [1] is a very popular and cheap (30\$) credit-card-sized single-board computer created in UK with the idea of teaching computer science and democratizing it in schools. The growth of Internet and Open Source allows technology to be available for everyone and cheap powerful motherboards are born. This is the case of the Raspberry project manufactured by Atmel®. The scientific field of image processing and the technologies associated are more and more present in the life of anyone by its numerous applications. This paper is devoted to show the capability of this device to process the images and videos and especially the classification of facial expressions. The specifications of RB are described in the first part, then the second part presents the image processing and pattern recognition algorithms utilized in the proposed method and finally a conclusion and future works will be given.

2 Small description of RP

Hereafter, the small size of the Raspberry Pi B used in this paper appears in figure 1 above the keyboard, the small camera Pi is shown on the upper right side. Raspberry Pi B has been preferred to Raspberry Pi A because it contains double more RAM useful for image processing.



Figure 1: The different devices used around the R.P. B, except the screen too big in the figure.

The different functions of RP appears on next figure 2:

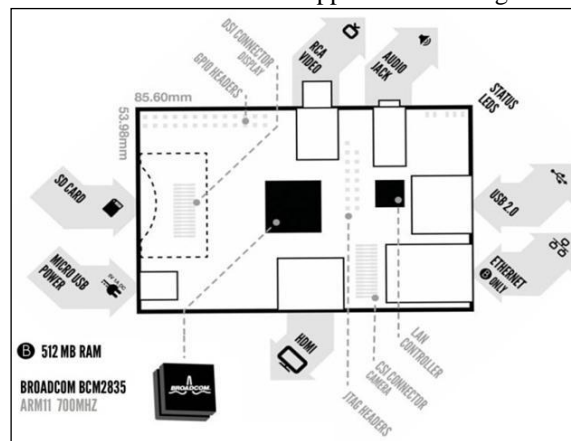


Figure 2. Functional parts of RP B

Most important elements to quantify the performances in image processing are presented in the following. It has a CPU ARM1176JZFS with 700MHz , GPU, 1Gpixels/s, 1.5Gtexel/s or 24 GFLOPS of general purpose compute and features a bunch of texture filtering and DMA infrastructure. The memory 512MB RAM is shared between CPU AND GPU. The camera CMOS with CSI can take images up to 5MP (2592*1944) and videos up to 1080p (1920*1080*30fps). The secondary memory is composed of a SD Card (8GB). All these specifications show the interest of the RP B for processing digital images for 30\$!. The Pi camera (Picam) has been paid 20^E, a web cam can be also used through USB input/output but it is more expensive. Picam is connected to the R.P. via a CSI port which permits to let one USB port free on the R.P. but has a drawback : it is not directly compatible, a specific library must be used during implementation. Unfortunately, the image processing functions of librairies such as OpenCV do not directly work with the PI cam, another library must be used: Raspicam, created by Emil Valkov, which creates a link between the Pi Camera's library and OpenCV. The operating system used is Raspbian linux. In summary, the software structure appears below:

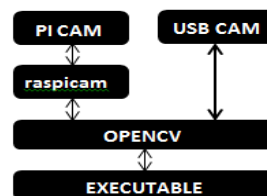


Figure 3. Software structure between camera Pi and openCV

The language C++ is utilized, because the other possible languages (python and processing) are much slower.

3 Proposed method

The first part of our work consists in detecting the face by surrounding it with a rectangle. The algorithm currently carried out is the Viola Jones one because it is efficient and permits real time face detection. It is well implemented in the openCV library. The 3 different steps of this technique are rapidly explained in the following. The integral image is performed for the feature extraction, Ada-Boost is carried out for face detection and finally the attentional cascade permits to the fast rejection of non face sub-windows. The integral image allows a very efficient computation of difference between the grey values of pixels belonging to different rectangles adapted to the shapes of eyes, nose and mouse. After this feature computation, a classification technique ada-boost founded on a combination of weak classifiers in order to detect faces in the input image. A cascade of ada-boost strong classifiers permits to remove many rectangular windows of non-faces and at the end of this pipeline face windows are provided. The results on our Raspberry Pi B of this Viola Jones method appears below.

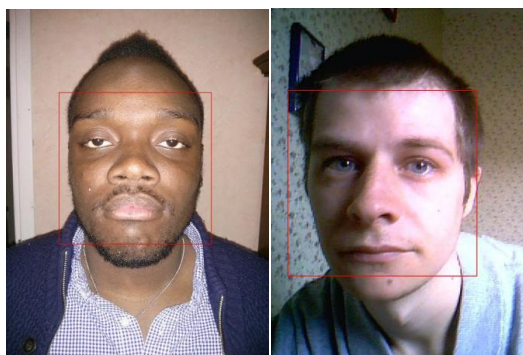


Figure 4. Rectangles around the faces

Then, the color images are converted into grey-level images because our experiments show it is sufficient for recognizing the normal, sad, smiling, laughing expressions of faces. As shown below, the rectangles around the mouth and eyes are automatically computed from locations of rectangles in the original images given by the Viola Jones procedure. The locations of rectangles around mouths and eyes are improved or adapted to each face by using projection on the y-axis inside these rectangles.

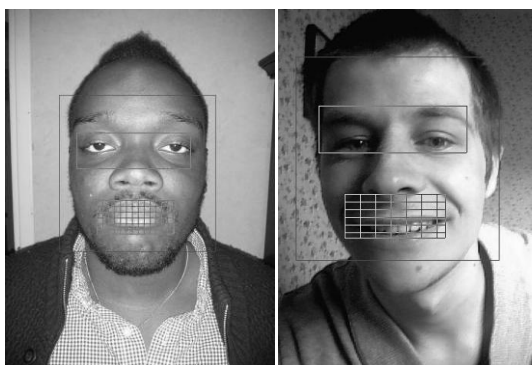


Figure 5. Rectangles around mouths and eyes

Principal Component Analysis (PCA) with eigen-faces and Linear Discriminant Analysis (LDA) with Fisher-faces have not been tried because the computations consume too much time but efficient. Furthermore, different methods using kernels have improved the results of these 2 methods during the last decade. This is the reason why Local Binary Patterns (LBP) [2] have been calculated on the grey-level images at pixels of mouth and eye rectangles. This technique which is enough robust to the lightning changes and is easy and quickly computed at the different pixels of the sampled pixels of mouth and eyes. LBP consists in taking any pixel in an image and its 8 neighbor pixels and deliver 1 if they are greater than the central pixel and 0 otherwise. If choosing the sequence of Freeman code, the LBP code of a pixel can be 10111101 which corresponds to a number between 0 and 255. Another measure called the contrast consists in making the difference of the 2 averages of grey values attached to the pixels with 1 and other ones with 0. In order to be more reliable to image noise, it is possible to use a fuzzy version of LBP. A more sophisticated version of LBP can be read in [2], it introduces the dynamics of the video and Volume LBP with right and left shifts of LBP codes and other good ideas.

In practice, different versions of LBP have been tested by varying the number of sampled pixels and the distance of neighbors to these pixels during the learning phase. Statistical parameters taking into account the LBP and spatial information have been computed and memorized during this off line stage and used during the operating phase. It permits to classify the facial expressions into normal, sad, smiling and laughing expressions. The results are good but can be improved by using points of interest linked to researches in neurosciences as quoted in [3] but it needs an additional computing step which is CPU consuming.

4 Conclusion

An automatic facial recognition technique has been described and proposed on very cheap computer and camera, it works well. Efforts have been performed in the choice and development of real time algorithms of image processing and pattern recognition adapted to low cost devices.

5 References

- [1] www.raspberrypi.com
- [2] G. Zhao, M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions", IEEE Trans. PAMI, 2007
- [3] M. Calvo et al, "Facial expression recognition in peripheral versus central vision: role of the eyes and mouth", Psychological Research, april 2013

Digital Image Processing in Surface Engineering Quality Control

E. Sheybani, S. Garcia, G. Javidi
Virginia State University, Petersburg, VA 23834

Abstract - Ability to measure the surface quality in real-time has many applications in manufacturing automation and product optimization, especially in processes in which the surface qualities such as roughness, grain size, thickness of coating, impurities size and distribution, hardness, and other mechanical properties are of importance. Surface analysis in manufacturing environments requires specialized filtering techniques. Due to the immense effect of rough environment and corruptive parameters, it is often impossible to evaluate the quality of a surface that has undergone various grades of processing. This research aims at providing image processing tools for comparison and assessment of a surface processed under different grades of a manufacturing process all the way up to optimal processing. The algorithm used is capable of performing this comparison analytically and quantitatively at a low computational cost (real-time) and high efficiency. The parameters used for comparison are the degree of blurriness and the amount of various types of noise associated with the surface image. Based on a heuristic analysis of these parameters the algorithm assesses the surface image and quantifies the quality of the image by characterizing important aspects of human visual quality. Extensive effort has been set forth to obtain real-world noise and blur conditions so that the various test cases presented could justify the validity of this approach well. The tests performed on the database of images produced valid results for the proposed algorithm consistently.

Keywords: DSP, Filter, Surface imaging, Wavelet, Processing

1 Introduction

This study starts by assigning a value to the visual quality of several images from the same object whose surface has undergone different grades of the same process. A digital image processing algorithm consisting of Wiener, Gaussian, un-sharp masking, and Multi-resolution wavelet filter banks is used to identify when an object is optimally processed. A delicate balance of these filters in the proposed algorithm is capable of recognizing the quality of an optimal surface (image) in comparison to unfinished versions of it. In doing so, we have identified some of the important parameters that affect the quality of a surface image and ways in which they can be measured quantitatively.

This research is aimed at accelerating research in theories, principles, and computational techniques for surface quality assessment. The results obtained indicate that the proposed algorithm can effectively assess the quality of any given surface from a wide range of variations in finish. Additionally, the proposed algorithm is fast, efficient, and robust and can be implemented in hardware for real-time applications.

In many advanced and automated industrial and manufacturing processes image processing algorithms (and hence surface images) are employed to analyze object surfaces and use the information obtained to improve the quality of the product such as finish, lighting, texture, color, placement, temperature, cracks, etc. [1],[2],[3]. One major disadvantage of these techniques is that collective environmental noise, speckles, and other artifacts from different sensors degrade the surface image quality in tasks such as surface pattern restoration, detection, recognition, and classification [4],[5]. While many techniques have been developed to limit the adverse effects of these parameters on surface image data, many of these methods suffer from a range of issues such as computational involvement of algorithms to suppression of useful information [6],[7]. Therefore, there is a great demand for a tool that could perform an accurate surface quality assessment. Since most surface defects in industrial environments look like clutter, noise, and/or phase/pixel shifts in imaging systems, we have based this proposed surface quality assessment algorithm on these parameters (noise and blurriness of surface image) [8],[9]. This is only a first step for a simple case of surface quality assessment (gray-scale, constant lighting surface images) and is aimed at answering some basic but important questions in surface quality assessment.

2 Conclusion

This effort has led to accelerated research in theories, principles, and computational techniques for surface quality assessment in image processing. The results obtained indicate that the proposed algorithm can effectively assess the quality of any given surface from a wide range of variations in finish. Furthermore, the algorithm can differentiate between a regular surface image corrupted with noise, blur and other clutter vs, one corrupted with artificial anomalies such as extra lighting, edges, etc. This further

validates functionality and accuracy of the proposed algorithm.

3 References

- [1] Maxwell, J. D.; Qu, Y.; Howell, J. R.; (2007). Full Field Temperature Measurement of Specular Wafers During Rapid Thermal Processing. *IEEE Transactions on Semiconductor Manufacturing*, Volume: 20 , Issue: 2, 2007 , Page(s): 137 – 142.
- [2] Kuo-Cheng Huang; Chun-Li Chang; Wen-Hong Wu; (2011). Novel Image Polarization Method for Measurement of Lens Decentration. *IEEE Transactions on Instrumentation and Measurement*, Volume: 60 , Issue: 5, 2011 , Page(s): 1845 – 1853.
- [3] Yuan Cheng; Jafari, M. A.; (2008). Vision-Based Online Process Control in Manufacturing Applications. *IEEE Transactions on Automation Science and Engineering*, Volume: 5 , Issue: 1, 2008 , Page(s): 140 – 153.
- [4] Du-Ming Tsai; Jie-Yu Luo; (2011). Mean Shift-Based Defect Detection in Multicrystalline Solar Wafer Surfaces. *IEEE Transactions on Industrial Informatics*, Volume: 7 , Issue: 1, 2011 , Page(s): 125 – 135.
- [5] Bernardini, F.; Martin, I.M.; Rushmeier, H.; (2001). High-quality texture reconstruction from multiple scans. *IEEE Transactions on Visualization and Computer Graphics*, Volume: 7 , Issue: 4, 2001 , Page(s): 318 – 332.
- [6] Pluim, J.P.W.; Maintz, J.B.A.; Viergever, M.A.; (2003). Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, Volume: 22 , Issue: 8, 2003 , Page(s): 986 – 1004.
- [7] Reed, S.; Ruiz, I.T.; Capus, C.; Petillot, Y.; (2006). The fusion of large scale classified side-scan sonar image mosaics. *IEEE Transactions on Image Processing*, Volume: 15 , Issue: 7, 2006 , Page(s): 2049 – 2060.
- [8] Sheikh, H.R.; Bovik, A.C.; (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, Volume: 15 , Issue: 2, 2006 , Page(s): 430 – 444.
- [9] Zhou Wang; Bovik, A.C.; (2002). A universal image quality index. *IEEE Signal Processing Letters*, Volume: 9 , Issue: 3, 2002 , Page(s): 81 – 84.

Video based Abnormal Behavior Analysis System for Surveillance

Jong-Gook Ko

Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
jgko@etri.re.kr

Abstract— For people safety and society security, fainting and abandonment detection are main issues in surveillance system recently. In this paper, we propose abnormal behavior analysis system that provide fainting and abandonment detection for video surveillance. Also we made used of human detection function for abnormal behavior analysis system to remove false alarm like fainting detection of non-human object and abandonment detection of human object. And we show GUI based human behavior analysis system that detects fainting and abandonment.

Keywords-Video Surveillance; Abnormal Behavior Detection, Fainting Detection, abandonment Detection;

I. INTRODUCTION

Intelligent video surveillance system is to monitor the activity of objects in a video. Visual surveillance has a wide variety of applications such as homeland security, crime prevention, and traffic control and so on. Intelligent video surveillance system should support not only basic object detection and tracking, but also interpret object abnormal behavior pattern finally. Fainting detection for safety and abandonment detection for crime prevention are main issues in surveillance system in recent years.

So far, many analysis researches are doing for Fainting and abandonment detection. For fainting detection, [1] analyze the bounding box coordinates representing the person in a single image. MHI and ellipse enclosed human body are used to detect falls [2]. The motion gradient and human shape features variation using the video are used in [3]. For abandonment detection, [4] used short-term and long-term blob split and merges for detect abandoned objects.

In this paper, we propose human behavior analysis system. It provides fainting and abandonment detection that reduce false alarm like fainting detection of non-human object and abandonment detection of human object. Proposed method has shown good performance in detecting of abnormal behaviors such as fainting and abandonment.

Section II describes system description, and we show test results in Section III. Finally, we conclude section IV.

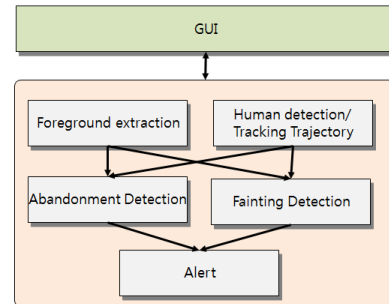


Figure 1. Overall system structure

II. SYSTEM DESCRIPTION

Fig.1 shows overall system structure. Abnormal behavior analysis system consists of foreground extraction, Human classification, abandonment detection, and fainting detection parts.

A. Foreground extraction and Tracking Trajectory

In this paper, we use codebook [5] to detect foreground object. Each foreground object's information such as location (x, y), size, and aspect ratio are gathered and maintained in every frame. Distance information between foreground objects is used for tracking trajectory of each object. That is, each object ID and location information is updated according to the rule of close distance in every frame.



Figure 2. Object detection and tracking

B. Human detection

For human detection, HOG features and SVM classification method are used. HOG features have been successfully used in pedestrian classification by Dalal [6]. It seems to be one of the best features for capturing edge and shape information, while being sensitive to noisy background edges and clutter. Each foreground object is checked whether human object or not by human

Mobile Visual Search Applications

Sung-kwan Je, Seungjae Lee and Weon-Geun Oh

Contents Protection Lab., Electronics and Telecommunications Research Institute, Daejeon, Korea

Abstract - In this paper, we implemented to MVS applications such as street searching service for navigation and visual searching service for something. When people are searching for somewhere or something, visual searching service is to let us know information what it is. The visual searching service is simple and an effective way to search and retrieval information and can help us search and find out something more intuitively.

Keywords: MVS, visual descriptor, retrieval information, street searching service, Visual Searching Service

1 Introduction

Developing mobile devices, it provides a convenience for user to easily search and retrieval information among huge of data has been getting an attention. But, a user wants to classify it, how to describe its contents in words which could be searched for. In current text-based search, we can obtain right information only if we know the exact name or related keywords. Otherwise, information can be accessed by image searching method of taking a picture. Image searching method is simple and an effective way to search and retrieval information. The MVS (Mobile Visual Search) technology is extracted the visual descriptor of the query image in mobile device and retrieved descriptor in database. For MVS, there are the following requirements [1]:

- Robust visual descriptor: the visual descriptor is robust against different lightening conditions and partial occlusion by other object.
- Reliable reference DB: the reference image are easily generated and updated to the DB
- Efficient searching structure: for piratical service with a large-scale reference DB, the efficient search structure is essential.

Recently, the MVS based service is launched as follows: In the Goggles of Google, Google+ supports visual similarity search, local search, product catalog has expanded service area [3]. Amazon was acquiring 'Snap Tell' for image visual search and developed own online shopping service. The shape and color based searching service is launched such as product search app 'Flow' and shoes search app 'Fabulous' [4]. Qualcomm also was acquiring 'kooaba' for image

recognition and launched such as MVS API based cloud recognition service 'Vuforia' [5]. Most existing MVS based searching services consist of three steps. In the first step, a user takes a picture of something to know information such as what it is, or where to buy, and the query image is send to the server. In the second step, a search engine retrieves the related image among the reference DBs. Finally, a user receives the information from the searching engine.

In this paper, we implemented to MVS applications of several situations such as outdoor, in the living room, and something curious object. In outdoor, street searching service is to let us know how to get somewhere. When a user watches a TV or finds something, visual searching service is to let us know information what it is.



Fig. 1. Street Searching Service for Navigation.

2 Applications

2.1 Street Searching Service for Navigation

When people are searching for somewhere or getting directions, we usually use a map or a map app based the GPS like Google map. The location based service is to let us know how to get somewhere using GPS to estimate the geographical position. But, it has some errors caused by a lot of things such as structural factors and the geometric errors.

In this paper, we implemented to MVS applications for street searching service. The people usually use a crossroad

sign, corner building or landmarks for searching. So, the buildings around crossroads are more appropriate for the image based localization. Our service scenario starts from crossroads as shown in Fig. 1. As the first step, a user takes a photo of the buildings around the crossroads. The query photo is transmitted to the searching server. In the second step, the user receives the location and he or she is asked to determine which direction is to be navigated. In the third step, the user looks around the selected direction with traditional map and multi-perspective panoramic street views. It can help us search and find out somewhere more intuitively.



Fig. 2. Visual Searching Service for Something using Leap motion device

2.2 Visual Searching Service for Something

When a user watches a TV or video in the living room, it wants to know something what it is such as cleaner with brand name and how to use. But, the content let us does not know what it is because of PPL (Product Placement). Usually, we can obtain right information only if we know the exact name or related keywords. Otherwise, information can be accessed by MVS of taking a picture using Leap motion device. The visual searching service is simple and an effective way to search and retrieval information as shown in Fig. 2. It can help us search and find out something more intuitively.

Another situation, when a user finds a something what it is such as cosmetics and how to use. Especially, a man does not know what it is because of too complication. The visual searching service is simple and an effective way to search and retrieval information as shown in Fig. 3. It can help us search and find out something more intuitively.

Increasing a user demand for searching of deformable object such as clothing, shoes, bag and so on, it will be examined by various deformable object and environmental conditions for more practical applications.



Fig. 3. Visual Searching Service for Something

Acknowledgment This work was supported by the IT R&D program of MSIP [R2012030111, Development of The Smart Mobile Search Technology based on UVD (Unified Visual Descriptor)]

3 References

[1] N11674, Compact Descriptors for Visual Search: Draft Call for Proposals, Requirements Subgroup, MPEG output document, Oct. 2010, Guangzhou, CN.

[2] Seungjae Lee et al.: Technology and Standardization Trend of Mobile Visual Search, Electronics and Telecommunications Trends, Vol.29, No.1, ETRI, 2014

[3] Goggles: [online]. <https://www.google.com/mobile/goggles>

[4] Flow: [online]. <https://flow.a9.com>

[5] Vuforia: [online]. <https://www.vuforia.com>

[6] IQ Engines: [online]. <http://www.crunchbase.com/company/iq-engines>

[7] Kooba: [online]. <http://www.kooba.com/>

[8] Moodstocks: [online]., <https://moodstocks.com/>

[9] N13951, White Paper on Compact Descriptors for Visual Search, ISO/IEC JTC1/SC29/WG1 MPEG, 2013.

[10] N12202, Compact Descriptors for Visual Search: Evaluation Framework, ISO/IEC JTC1/SC29/WG1 MPEG, 2011.

Global Localization of Mobile Robot using an Omni-Directional Camera

Seung-Hun Kim

Intelligent Robotics Research Center
 Korea Electronics Technology Institute
 Bucheon, Gyeonggi-do, Korea
 ksh1018@keti.re.kr

Ju-Hong Park, Il-Kyun Jung

Intelligent Robotics Research Center
 Korea Electronics Technology Institute
 Bucheon, Gyeonggi-do, Korea
 juhong, mickey3d@keti.re.kr

Abstract—In this paper, we propose a method for global localization using an omni-directional camera. A robot position and angle are estimated by correlation coefficient between topological node-map images and input images. Near-node has the largest correlation coefficient in topological map images. The calculated correlation coefficient makes the mixtures of Gaussians density map. The highest value of mixtures of Gaussians density map is the estimated robot position. The angle of the robot is estimated using KLT method between near-node image and input image. The experiment is practiced in the lobby at Bucheon RoboPark building. The proposed algorithm is proved by the experimental results.

Keywords : Mobile Service Robot, Global Localization, Omni-Directional Camera, Correlation Coefficient Value, KLT

I. INTRODUCTION

Localization is essential technology when a service robot moves to the destination. A service robot has laser scanner, ultrasonic sensors, cameras, and several sensors to obtain environment information for localization. A camera is low cost and can be obtained various information. For that reason, localization studies have been actively performed using a camera[1][2]. However, a camera is difficult to obtain sufficient information by a small FOV(Field Of View). In this paper, we solve the problem by using a omni-directional camera. A robot position and angle are estimated by correlation coefficient between topological node-map image and input image. The robot installed an omni-directional camera module is shown in Figure 1.

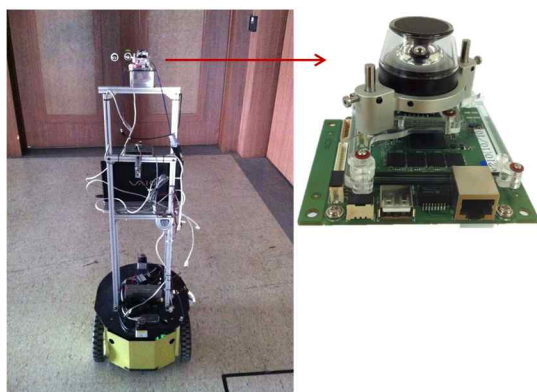
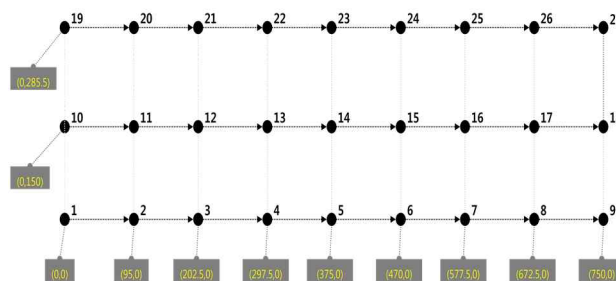


Figure 1. Mobile robot system with the omni-directional camera modules

II. CORRELEATION COEFFICIENT LOCALIZATION

We need a topological node-map with reference image of each node for global localization. Each node has a world coordinate orientation and an omni-directional image. Topological node map and the node map image are shown in Figure 2. Correlation Coefficient Value(CCV) is obtained between topological node-map image and input image. Correlation coefficients are computed using FFT(Fast Fourier Transform). We change an omni-directional circle image into a rectangle panoramic image for obtaining CCV.



(a) Topological node map



(b) Node map image

Figure 2. Topological node

Proceeds as follows.

- A. Unwarp input image.
- B. Calculate CCV between topological node map image and input image.
- C. Determine near-node by the largest CCV in topological map images.

- D. Make weighted mixtures of Gaussians density map by CCV.
- E. Estimate the robot position by the highest value of mixtures of Gaussians density map
- F. Rotate input image by the maximum value position of Near-node CCV.
- G. Find the corner points feature in near-node image and rotated input image.
- H. Calculate the robot angle by calculating the amount of rotation between two images using KLT.

A. Unwarping Image

Omni-directional circle image transform to panoramic image using Concentric circle approximation unwarp method.

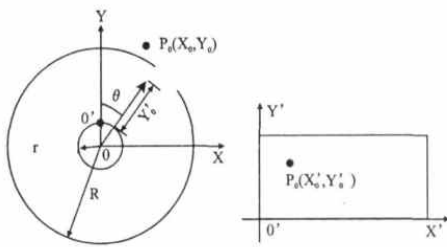


Figure 3. Concentric circle approximation unwarp method

r: inside diameter, R: outside diameter

$$\begin{aligned} X_0 &= (r + X'_0) \times \sin H \\ Y_0 &= (r + Y'_0) \times \cos H \end{aligned} \quad (1)$$

where, $H = X'_0 / (r + Y'_0)$



(a) Original image



(b) Result image

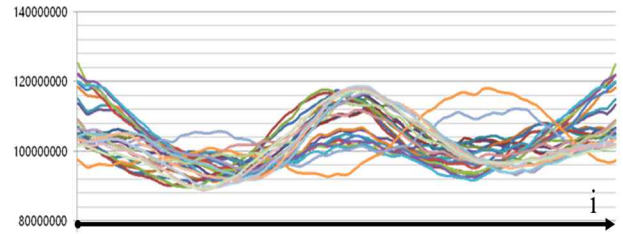
Figure 4. Unwarping

B. Correlation Coefficient Localization

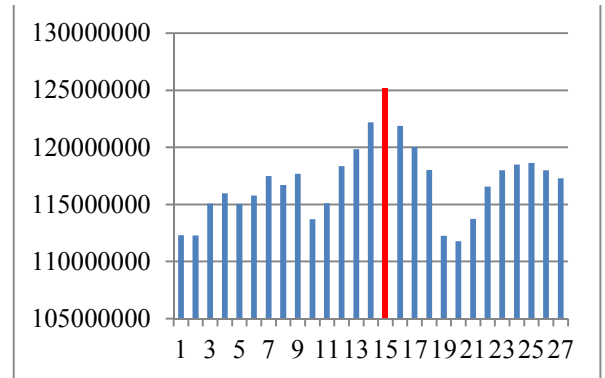
Correlation coefficients are computed using FFT. Correlation coefficient value equation is below.

$$Corr(x, y)_i = F^{-1} \{F(x)F^*(y)\} \quad (2)$$

Near-node was determined by the largest CCV in topological map image(Figure 5). And the maximum value position of the correlation coefficient between near-node image and input image is the rotate angle value.(Figure 6)



(a) CCV graph for all node images



(a) Determine near-node

Figure 5. Determine near-node by maximum CCV

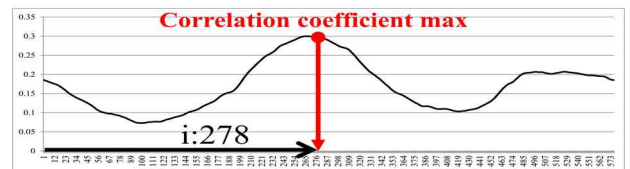


Figure 6. Rotation value

Mixtures of Gaussians density map is made by each node CCV weighted. The highest value of mixtures of Gaussians density map is the global location of the robot.

$$p(x | \lambda) = \sum_{i=1}^M w_i g(u_i, \sigma_i) \quad (3)$$

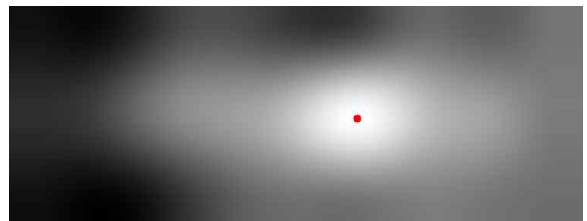
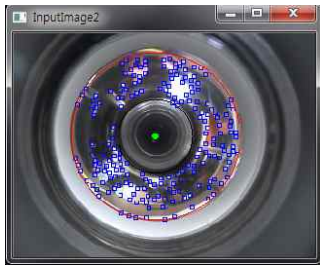


Figure 7. Determine robot position in mixtures of Gaussians density map

C. Robot Angle Estimateion using Harris Corner and KLT

We use feature matching between near-node image and input image for the rotation angle of the robot. The corner

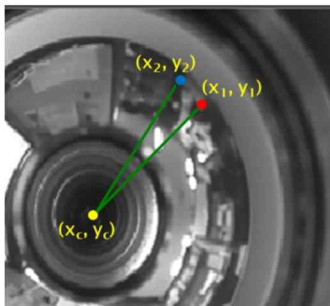
feature points in near-node image and rotated input image are obtained by Harris corner. The rotation angle of the robot is calculated as the amount of rotation between two images using KLT.



(a) Harris corner



(b) Points matching between near-node and input image using KLT



(c) Rotation angle estimation by 1 point

Figure 8. Robot angle estimation

The rotation angle of the robot is estimated to be the following equation 4. (P1 is the point of the near-node image. P2 is point of the rotated input image. They are the same point.)

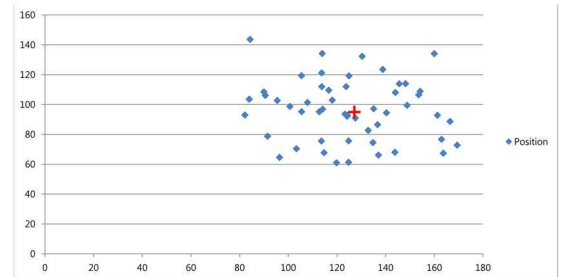
$$angle = a \cos\left(\frac{p1 \bullet p2}{|p1| |p2|}\right) \quad (3)$$

III. EXPERIMENTAL RESULTS

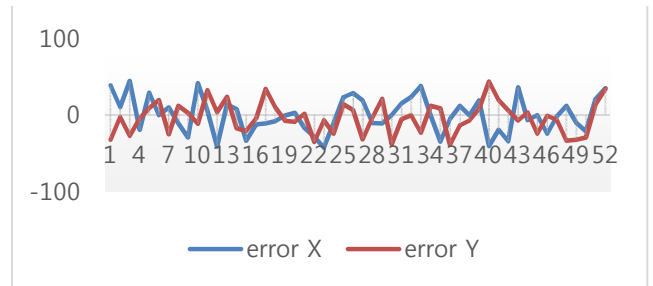
To verify the performance of the proposed method, we had constructed node map, as shown in Figure 9. The total number of nodes is 27 and node spacing is 1m. The robot is located at the known ground truth pose. The computing time is 390.4 msec. The average of the distance error(mean error) is 30 cm and the angular error is 0.082 degrees.



Figure 9. Node map



(a) Result of position estimation.

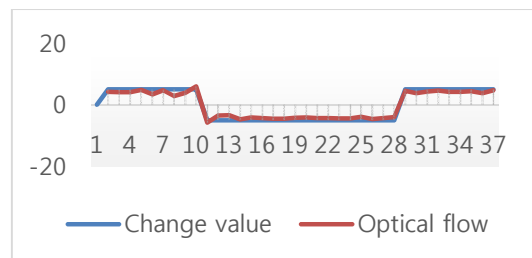


(b) Position estimation error

Figure 10. Result of global localization



(a) Rotation experiment.



(b) Rotation estimation error

Figure 11. Result of rotation estimation

IV. CONCLUSIONS

This paper proposed the global localization using correlation coefficient and KLT method. The developed module is a single omni-directional camera looking at the around of the robot and inexpensive and easy-to-get everywhere. Experiment was tested in real environment and global localization method is proposed without initial information. Correlation coefficient method is fast compared to other invariant feature point matching methods.

REFERENCES

- [1] S. Se, D. G. Lowe, and J. J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Transactionson Robotics*, vol. 21, no. 3, pp. 364–375, 2005.
- [2]
- [3] Reynolds, D. A.: *Gaussian Mixture Models*, *Encyclopedia of Biometric Recognition*, Springer (2008).
- [4] C. Harris and M. Stephens (1988). "A combined corner and edge detector". *Proceedings of the 4th Alvey Vision Conference*. pp. 147–151.
- [5] C. Tomasi and T. Kanade. *Detection and tracking of point features*. Technical report CMU-CS-91-132, Carnegie Mellon University, 1991.

SESSION

LATE BREAKING PAPERS AND POSITION PAPERS: IMAGE PROCESSING, COMPUTER VISION, AND PATTERN RECOGNITION

Chair(s)

Prof. Hamid R. Arabnia

Runway and Horizon Detection through Fusion of Enhanced Vision System and Synthetic Vision System Images

Ahmed F. Fadhil, Raghuveer Kanneganti, and Lalit Gupta

Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA

Abstract - A novel procedure is developed to accurately detect runways and horizons and also enhance runway surrounding areas by fusing enhanced vision system (EVS) and synthetic vision system (SVS) images of the runway while an aircraft is landing. Because the EVS and SVS frames are not aligned, a registration step is introduced to align the EVS and SVS images prior to fusion. The most notable feature of the registration procedure is that it is guided by the information extracted from the weather-invariant SVS images. A fusion rule based on combining DWT sub-bands is implemented and evaluated. The resulting procedure is tested on real EVS-SVS image pairs and also on image pairs containing simulated EVS images with varying levels of turbulence. The subjective and objective evaluations reveal that runways and horizons can be detected accurately even in poor visibility conditions. Another notable feature is that the entire procedure is autonomous throughout the landing sequence irrespective of the weather conditions. That is, the variable parameters are determined automatically from the image frames during operation. Given the excellent fusion results and the autonomous feature, it can be concluded that the fusion procedure developed is quite promising for incorporation into head-up displays (HUDs) to assist pilots in safely landing aircrafts in varying weather conditions.

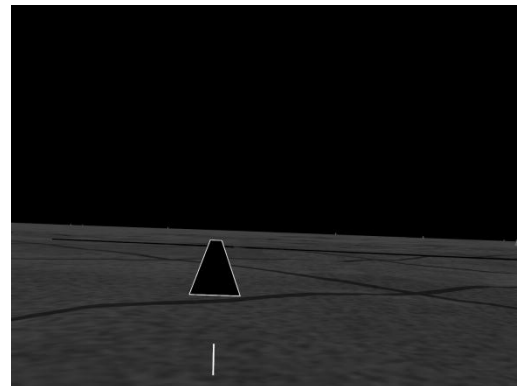
Keywords: Runway detection, Image fusion, Image registration, Wavelet transform, Hough transform

1 Introduction

The precise detection of runways is crucial for safely landing aircrafts because more than half of the accidents occur during the final approach and landing stages [1]. Instrument landing systems which provide precise landing guidance are not available at all airports. The challenge, therefore, is to assist pilots using visual flight landing rules to detect runways accurately in varying weather conditions. The runway and horizon detection and enhancement approach introduced in this study is based on exploiting information from enhanced vision system (EVS) and synthetic vision system (SVS) images of the runways. The goal is to generate image frames that contain enhanced runway and surrounding information by fusing the EVS and SVS images. The resulting image frames can be incorporated into head-up displays (HUDs) to assist pilots in safely landing aircrafts.



(a)



(b)

Fig. 1 An example of (a) EVS and (b) SVS frames

Figure 1 shows an example of corresponding EVS and SVS images of a runway. The EVS image is an infra-red image of the runway and the SVS image is a GPS generated image of the runway. The EVS image can be used by pilots during landing but the quality of the EVS image is affected by adverse weather conditions. Unlike the EVS image, the SVS image is not affected by the weather. However, the SVS image is not a “real” image of the runway and cannot be used solely to help the pilot safely land the aircraft. The approach developed in this paper is to exploit the weather-invariant SVS image information to accurately detect the runway in the weather-dependent EVS image and to generate an EVS-SVS composite image which contains information from both images. A fusion rule is developed to combine the EVS and

SVS images and evaluated in the presence of varying levels of atmospheric turbulence. Because the EVS and SVS images are not aligned, the images are registered using runway and horizon features prior to fusion.

It must be emphasized that the goal in this study is to not only accurately detect the runway in a sequence of frames through registration but to also enhance the information surrounding the runway by fusing the EVS and SVS images. Because the most critical need for accurate runway detection is during the landing phase, the focus will be on registering and fusing the EVS and SVS images only when the aircraft is close to the runway. The procedure is made autonomous from frame-to-frame by deriving parameters from the SVS image. The performance of the fusion methodology is evaluated on a data set consisting of 1350 pairs of EVS and SVS frames of the runway acquired while an aircraft was landing. Furthermore, additional subjective and objective evaluations are conducted using simulated EVS images with varying levels of atmospheric turbulence. Related work include noteworthy studies which focuses on using synthetic vision data to accurately locate the position of the runway and then detect moving objects on the runway [2], fusing real and virtual information to enhance airport scenes [3], and integrating EVS and SVS data to improve visibility in adverse atmospheric conditions [4]. The work described in this paper differs from those in references [2,3,4] because the primary focus is on both runway and horizon detection as well as the fusion of EVS and SVS images to provide improved visual information of the runway scene. Furthermore, the formulations of the steps to detect runways and horizons, to fuse images, and to evaluate the performance subjectively and objectively, differ from those described in [2,3,4].

2 EVS and SVS Image Registration

Because the goal of this study is to fuse the runways as well as the surrounding areas, features for registration are derived from the runways and the horizons. The features selected for registration are the runway corners and the horizon end-points. The runway quadrilateral is composed of two long line segments and two shorter line segments. The runway corners can, therefore, be determined by detecting the end-points of the two long line segments. The horizon end-points can be determined from the line segment corresponding to the horizon. The first step, therefore, is to detect line segments in the EVS and SVS images and select those line segments that correspond to the two longer runway segments and the horizon line. The EVS and SVS frames are represented by $f_E(x, y)$ and $f_S(x, y)$, respectively. Because the SVS image quality is good and is unaffected by the weather conditions, the runway and horizon are first detected in the SVS image. The SVS runway and horizon information are used to estimate the runway and horizon in the corresponding EVS image.

2.1 SVS runway and horizon detection

A careful examination of the SVS image frames reveal that the following useful information (i) the image is a simple gray-scale image, (ii) the runway is outlined by a bright boundary, and (iii) the horizon is a distinct boundary between the all-dark sky and the brighter non-sky regions. In the first step, the runway is detected by segmenting $f_S(x, y)$ according to

$$g_S(x, y) = \begin{cases} 1, & \text{if } f_S(x, y) > (\delta_S) \text{Max}[f_S(x, y)] \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The factor $\delta_S, 0 < \delta_S < 1$, is determined empirically so that the resulting binary image $g_S(x, y)$ contains only the bright runway border. The four corner points of the runway are selected as runway registration control points in the SVS image. The end-points of the horizon, which will serve as horizon registration control points, are found by detecting the transition point from the dark pixels to the brighter pixels in the first and last columns of the image. The line joining the end-points defines the horizon in the SVS image. Figure 2(a) shows, in blue, the runway and horizon extracted from the SVS image in Figure 1 using $\delta_S = 0.5$ which was found to give good segmentation results across all 1350 SVS frames. If $g_S(x, y)$ is the binary image of the runway and horizon, the angles of the two long runway line segments and the horizon line can be found from the co-linearity Hough transform $S(\rho, \theta)$ of $g_S(x, y)$. Let $\{S(\rho, \theta_1)\}$ be the Hough transform accumulator cell with the highest count and let $\{S(\rho, \theta_2)\}$, and $\{S(\rho, \theta_3)\}$ be the cells with the next two highest counts. Then, the horizon angle is given by θ_1 and the runway angles are given by θ_2 and θ_3 . The angle parameters will be used to determine the runway lines and horizon in the EVS image.

2.2 EVS runway and horizon detection

Unlike the SVS image, the EVS image can be relatively complex. Consequently, the runway cannot be determined directly through segmentation. Nor can the horizon be detected using the method developed for the SVS image. Moreover, the EVS frames are bound to be degraded with noise. In order to decrease the effects of noise, the EVS frames are filtered using a Weiner filter [5]. In the frequency domain, the EVS filtered image is given by

$$\widehat{F}_E(u, v) = \left[\frac{1}{H(u, v) |H(u, v)|^2 + K} \right] F_E(u, v) \quad (2)$$

where $H(u, v)$ is the degradation function, $F_E(u, v)$ is the Fourier transform of $f_E(x, y)$, and K is a specified constant. The main degradation is assumed to be atmospheric turbulence, therefore, the function

$$H(u, v) = e^{-k(u^2+v^2)^{5/6}} \quad (3)$$

which is often used to model turbulence [5], is selected. The constant k can be adjusted according to the amount of turbulence. The runway and horizon in the EVS image $\hat{f}_E(x, y)$ are detected using information extracted from the SVS image. The SVS runway and horizon serve as initial approximations for the EVS runway and horizon, respectively. In order to detect the runway in the EVS image, let $\hat{r}_E(x, y)$ be a rectangular region encompassing the approximated runway. The two long runway sides are within the diagonal and vertical orientations when the aircraft approaches the runway. Therefore, $\hat{r}_E(x, y)$ is converted into a region $\tilde{r}_E(x, y)$ containing vertical and approximately vertical lines by applying a (3×3) vertical line detection mask [19]. The region $\tilde{r}_E(x, y)$ is converted into a binary region $g_E(x, y)$ using the following segmentation rule:

$$g_E(x, y) = \begin{cases} 1, & \text{if } \hat{r}_E(x, y) > (\delta_E) \text{Max}[\hat{r}_E(x, y)] \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where, δ_E is determined empirically so that the resulting binary image contains only the bright line segments and removes the lower-intensity line segments. The Hough transform $E(\rho, \theta)$ of $g_E(x, y)$ is computed and the pixels contributing to accumulator cells $\{E(\rho, \theta_2 \pm \alpha)\}$ and $\{E(\rho, \theta_3 \pm \alpha)\}$ are selected to determine line segments that have approximately the same orientations as the SVS runway. The two largest line segments satisfying $\{E(\rho, \theta_2 \pm \alpha)\}$ and $\{E(\rho, \theta_3 \pm \alpha)\}$ are selected as the runway line segments in the EVS image. The parameter α is included to account for the fact that runway is not perfectly aligned in the SVS and EVS images. The end-points of these two line segments give the runway registration control points in the EVS image. Moreover, the lines connecting the runway control points define the estimated runway in the EVS image. The estimated runway, which tends to compactly enclose the actual runway, is superimposed onto the EVS image using an intensity equal to 255. In a similar manner, the horizon in the EVS image is estimated by using the SVS horizon as an initial approximation and finding the dominant line within $(\theta_1 \pm \alpha_1)$ from the Hough transform in a band encompassing the initial approximation. The two-end points of the horizon give the horizon control points in the EVS image. The line joining the two end-points is superimposed on the EVS image using an intensity equal to 255.

2.3 Horizon and runway registration

Figure 2(a) which shows the result of superimposing the SVS (blue) and EVS (white) runways and horizons. It is clear that runways and horizons are not aligned and, therefore, have to be registered prior to fusion. A two-step registration procedure is developed in which the two images are first globally aligned based on horizon registration and then locally aligning the runways in the horizon-aligned images. Because the key information for landing is in the real EVS image, the SVS image is registered to the EVS image. The horizon corner points, top image corner points, and the bottom image

corner points are the six pairs of control points selected for the horizon based registration. The runway corner points and the corner points of the runway encompassing rectangle $r(x, y)$ are selected as the control points for registering the runways. For both steps, the projective transformation is applied to register the images and the results are shown in Figure 2(b). Note that the blue and white lines are perfectly aligned and almost appear as single blue lines.

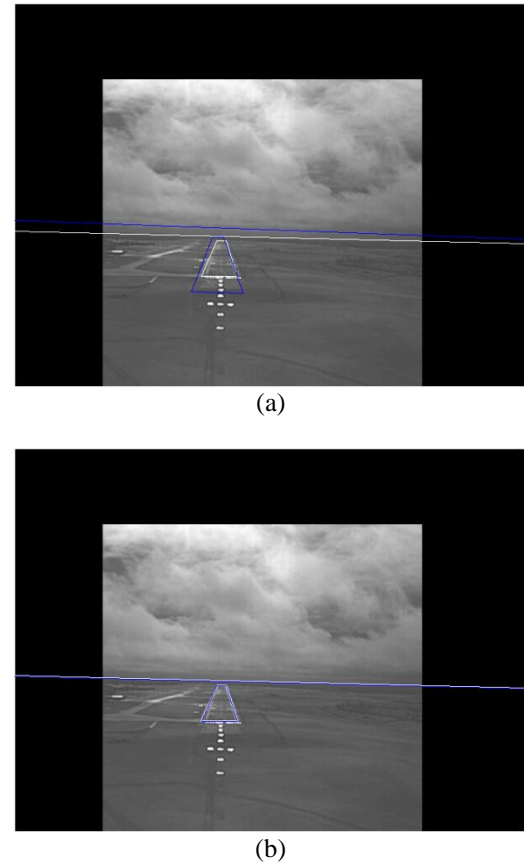


Fig. 2 Superimposed SVS horizon and runway onto the EVS image with (a) no registration (b) with registration

3 EVS and SVS Image Fusion

Image fusion is the process of combining two or more images in such a way that information from both images is preserved [6]. The images can be fused directly in the spatial domain [7] or in a transform domain such as the wavelet domain [6]. Although quite simple, the spatial domain rules are limited to the global application of operations such as pixel averaging or maximum selection [8, 9]. Conversely, the wavelet domain rules offer greater flexibility in developing fusion rules [7-12]. In this study equal weightage is not assumed because the EVS image is more important than the “supplementary” information in the SVS image. The goal here is to generate an image $w_{ES}(x, y)$ which is obtained by fusing the Weiner filtered EVS image $w_E(x, y)$ and the registered SVS image $w_S(x, y)$. If the dimension of $w_E(x, y)$ is

assumed to be $M \times N$, the DWT of $w_E(x, y)$ can be written as

$$W_E^a(m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} w_E(x, y) \varphi_{m,n}(x, y) \quad (5)$$

$$W_E^i(m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} w_E(x, y) \psi_{m,n}^i(x, y), \quad i = \{h, v, d\}$$

where $\varphi_{m,n}(x, y)$ and $\psi_{m,n}^i(x, y)$ are the scaled and translated Haar basis functions. The corresponding inverse wavelet transform is then given by:

$$w_E(x, y) = \frac{1}{\sqrt{MN}} \left[\sum_{m=0}^{M/2-1} \sum_{n=0}^{N/2-1} W_E^a(m, n) \varphi_{m,n}(x, y) + \sum_{i=h,v,d} \sum_{m=0}^{M/2-1} \sum_{n=0}^{N/2-1} W_E^i(m, n) \psi_{m,n}^i(x, y) \right]. \quad (6)$$

$W_E^a(m, n)$, $W_E^v(m, n)$, $W_E^h(m, n)$, and $W_E^d(m, n)$ are the four $(M/2) \times (N/2)$ sub-bands of the DWT of $w_E(x, y)$. These sub-bands are the approximation, vertical detail, horizontal detail, and diagonal detail sub-bands of $w_E(x, y)$. Similarly, $W_S^a(m, n)$, $W_S^v(m, n)$, $W_S^h(m, n)$, and $W_S^d(m, n)$ represent the sub-bands of the DWT of $w(x, y)$. Using the DWT sub-bands, the images are fused according to the following rule:

$$W_{ES}^a(m, n) = W_E^a(m, n) \quad W_{ES}^v(m, n) = \text{MAX}[W_E^v(m, n), W_S^v(m, n)]$$

$$W_{ES}^h(m, n) = \text{MAX}[W_E^h(m, n), W_S^h(m, n)]$$

$$W_{ES}^d(m, n) = \text{MAX}[W_E^d(m, n), W_S^d(m, n)] \quad (7)$$

where, $W_{ES}^a(m, n)$, $W_{ES}^v(m, n)$, $W_{ES}^h(m, n)$, and $W_{ES}^d(m, n)$ are the sub-bands of the EVS-SVS fused image. The fused image $w_{ES}(x, y)$ is obtained from the inverse wavelet transform after application of the selection rule. Note that most often, images are fused using the maximum, average, or mixed rules [7-12]. The above modified rule is introduced to give more weight to the EVS image by preserving the EVS information in the approximation band and can be regarded as a modification of the maximum and mixed rules.

4 Registration and Fusion Experiments

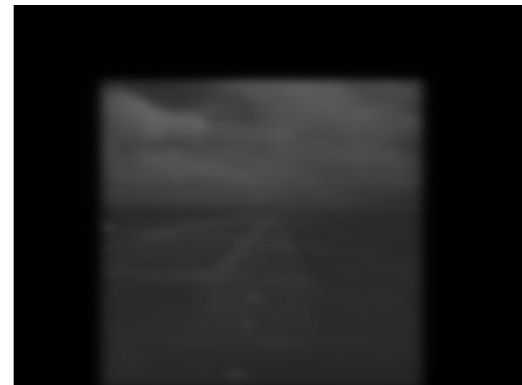
The data set, provided by Rockwell-Collins, consists of 1350 EVS and SVS image pairs acquired from the infra-red sensor on an aircraft and a satellite, respectively. The dimensions of EVS and SVS frames are 1050×1400 and the frame rate was 6 fps. The following values were used for the parameters: $\delta_S = 0.5$, $k = 0.0025$, $K = 10^{-8}$, $\delta_E = 0.6$, $\alpha = 30^\circ$, $\alpha_1 = 25^\circ$. After setting the parameters, the entire sequence of frame pairs were processed autonomously without any intervention.

Because it is not practical to show the results for all 1350 frame pairs, only a few select examples are shown to conduct subjective evaluations. Also, the real data set does not cover varying weather conditions. In order to conduct more detailed subjective and objective evaluations, data sets containing

various levels of atmospheric turbulence in the EVS frames are generated using the same degradation model $H(u, v)$ used to filter the EVS image. Two sets of EVS frames were generated to simulate intermediate level turbulence using $k=0.001$, and severe level turbulence using $k=0.0025$ in each frame. Examples of degraded EVS frames are shown in Figure 3. The registration performance can thus be evaluated objectively by the root-mean-square (rms) error between the manually detected runway corner points and horizon points in the EVS image, selected using mouse-clicks, and the corner points detected by the registration procedure.



(a)



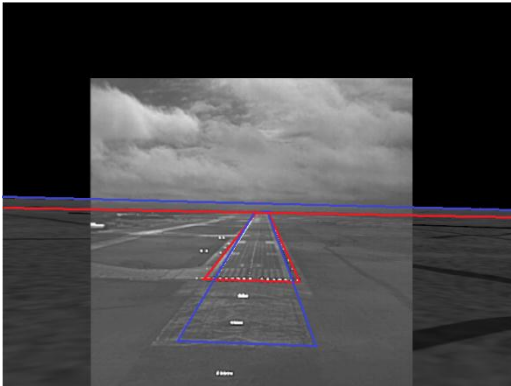
(b)

Fig. 3(a) EVS image with intermediate turbulence (b) EVS images with severe turbulence

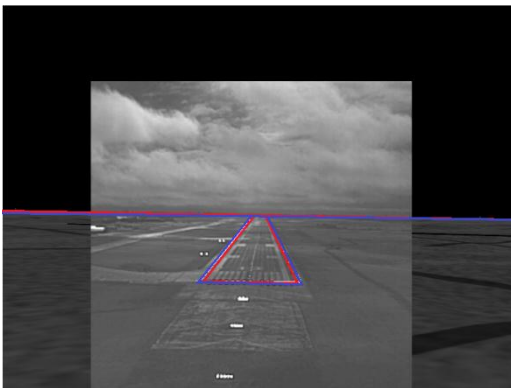
Figure 4 shows examples of fusing the EVS and SVS images directly without prior registration. The runways and horizons in the SVS and EVS images are shown in blue and red, respectively, to facilitate visual analysis. Observe the following problems in Figure 4(a) (i) the runways and horizons are not aligned and (ii) the SVS runway is much larger than the EVS runway. Figure 4(b) shows the fusion results of the same images after they have been registered. Observe that the runways and horizons are aligned quite well.

Examples of fusing pairs of registered EVS and SVS are shown in Figures 5 for the original (no-turbulence assumption) and severe turbulence cases, respectively. The most important results are for the severe turbulence case

because there is not much need for runway and horizon detection when there is little or no atmospheric turbulence. The no-turbulence results are included to for comparison purposes. The result in Figure 5(b) is quite impressive given that the input was the image in Figure 3(b). Observe that the important EVS information is preserved while also fusing the lesser important SVS information. Most importantly, the runway details are also clear which can also be useful for detecting obstacles on runways.



(a)



(b)

Fig. 4 Examples EVS-SVS image fusion (a) without prior registration (b) after registration



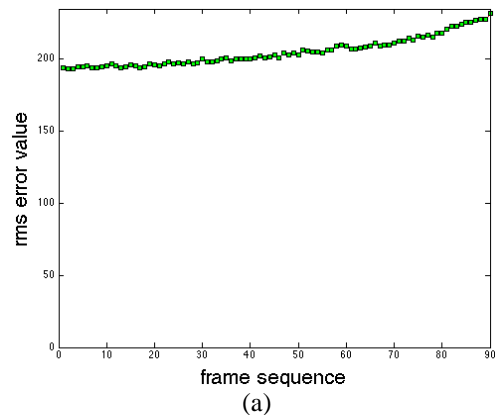
(a)



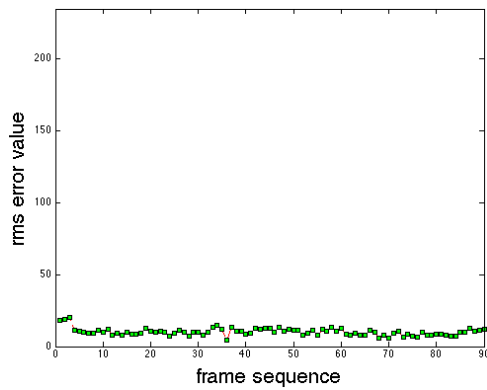
(b)

Fig. 5 Fusion in (a) no-turbulence (b) in severe turbulence

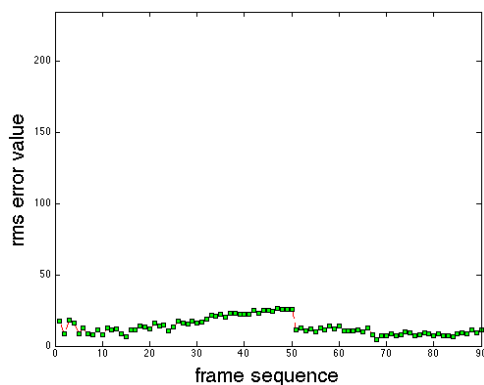
The next set of results are included to evaluate the registration performance by evaluating the rms distance error between the locations of the manually detected runway and horizon corner points and locations of the registered runway and horizon corner points. Figure 6 summarizes the rms errors obtained from the original data set without registration, the original data set with registration, and the severe turbulence degraded data sets with registration. The results cover frames 1101 to 1190. The frames are ordered such that the frame number increases as the distance between the aircraft and runway decreases. The original data set without registration is included to compare the registration performance. The following points should be noted regarding the results presented in Figure 6: (i) when no registration is employed, the rms errors are relatively high. Moreover, the rms errors increase when the distance between the aircraft and the runway decreases. This increase is clearly undesirable because it is even more critical to detect the runway accurately as the aircraft gets closer to the runway; (ii) The rms errors are smaller when registration is employed. Most importantly, the errors do not increase when the aircraft gets closer to the runway; (iii) The error trends across the registered results are quite similar. Furthermore, the increase in the rms errors in the presence of severe turbulence is marginal; (iv) The various parameters were exactly the same for the four sets of results. Note that there was no attempt made to adjust the parameters to accommodate the turbulence in Figure 6(c).



(a)



(b)



(c)

Fig. 6 Corner and horizon registration rms errors (a) original data set with no registration (b) original data set with registration (c) data set with severe turbulence.

5 Conclusions

A novel procedure was developed to accurately detect runways and horizons and also enhance surrounding runway areas by fusing EVS and SVS images. The primary focus was on fusing EVS and SVS images of the runway while an aircraft was in the final stages of landing. A registration procedure was introduced to align the EVS and SVS images prior to fusion. The most notable feature of the registration procedure was that it was guided by the information extracted from the weather-invariant SVS images. A fusion rule based on combining DWT sub-bands was implemented and evaluated. The resulting procedure was tested on real EVS-SVS image pairs and also on image pairs containing simulated EVS images with varying levels of turbulence. The subjective and objective evaluations revealed that the runways could be detected accurately even in poor visibility conditions due to severe levels of atmospheric turbulence. Another notable feature is that the entire procedure is autonomous throughout the landing sequence irrespective of the weather conditions. That is, the fixed parameters are set initially and the variable parameters are determined automatically from the image

frames during operation. Given the excellent fusion results and the autonomous feature, it can be concluded that the fusion procedure developed is quite promising for incorporation into head-up displays (HUDs) to assist pilots in safely landing aircrafts in varying weather conditions. Furthermore, the procedure developed can be easily modified to fuse image pairs in different applications. The observations and results can also serve as a guide for selecting different fusion rules for a given application.

6 Acknowledgements

This research has been supported by the NSF I/UCRC for Embedded Systems at SIUC, and funded in part by the National Science Foundation under Grant No. 0856039. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7 References

- [1] Boeing, "Statistical summary of commercial jet airplane accidents, worldwide operations 1998-2012," 2013.
- [2] R. Hamza, et al., "Runway positioning and moving object detection prior to landing," in *Augmented Vision Perception in Infrared: Algorithms and Applied Systems*, London, United Kingdom: Springer, 2009.
- [3] P. Cheng, et al., "A fusion of real and virtual information for aiding aircraft pilotage in low visibility," *Journal of Computers*, vol. 8, no. 4, pp. 874-877, 2013.
- [4] N. S. Kumar, et al., "Integrated enhanced and synthetic vision system for transport aircraft," *Defence Science Journal*, vol. 63, no. 2, pp. 157-163, 2013.
- [5] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," 3rd ed., New Jersey: Pearson Prentice Hall, 2008.
- [6] S. Nikolov, et al., "Wavelets for image fusion," in *Wavelets in Signal and Image Analysis: From Theory to Practice*, Dordrecht, The Netherlands: Kluwer Academic Publishers, 2001.
- [7] S. Li and B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," *Image and Vision Computing*, vol. 26, no. 7, pp. 971-979, 2008.
- [8] W. F., Jr., Herrington, B.K.P. Horn, and I. Masaki, "Application of the discrete Haar wavelet transform to image fusion for night time driving," in *Proceedings of IEEE Intelligent Vehicles Symposium*, pp. 273-277, June 6-8, 2005.

[9] K. Rani and R. Sharma, "Study of Different Image fusion Algorithm," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 5, pp. 288-291, 2013.

[10] D. Tseng, Y. L. Chen, and M. S. C. Liu, "Wavelet-based multispectral image fusion," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. 4, Sydney, NSW, pp. 1956-1958 July 9-13, 2001

[11] H. Zheng, et al., "Study on the Optimal Parameters of Image Fusion Based on Wavelet Transform," *Journal of Computational Information*, vol. 6, no. 1, pp. 131-137, 2010.

[12] S. M. M. Rahman, M. O. Ahmad, and M. N. S. Swamy, "Contrast-based fusion of noisy images using discrete wavelet transform," *IET Image Processing Journal*, vol. 4, no. 5, pp. 374-384, 2010.

Solving Optimization Problems That Employ Structural Similarity As The Fidelity Measure

Daniel Otero and Edward R. Vrscay

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada

Abstract—Many tasks in image processing are carried out by solving appropriate optimization problems. As is well known, the square of the Euclidian distance is widely used as a fitting term, even though it has been shown not to be the best choice in terms of quantifying visual quality. To overcome this problem, a number of papers have examined the use of the Structural Similarity Index Measure (SSIM) as a fidelity term. In this paper, we propose a general framework for solving optimization problems in which the SSIM is employed as a fidelity measure. Within the context of quasi-convex optimization, an algorithm is also introduced in order to solve such optimization problems.

Keywords: Structural similarity, ℓ_1 norm constrained optimization, total variation, quasi-convex optimization, deblurring, zooming

1. Introduction

Over the years, in the field of Image Quality Assessment (IQA), many metrics have been proposed to model the Human Visual System (HVS), including the Mean Opinion Score (MOS), the Universal Quality Index (UQI) [1], the Perception-based Measure (PBM) [2], among others. More recently, Wang *et al.* introduced in [3] the Structural Similarity Index Measure (SSIM), a visual metric that has been the focus of considerable research [4], [5], [6], [7], [8]. As opposed to error-based metrics such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), SSIM relies on the assumption that the HVS has evolved to perceive errors as changes in structural information, hence quantifying these changes should correlate better with our perception of what we consider good visually. Indeed, in [3] and other works, it has been shown that the SSIM outperforms other measures of visual quality, including MSE.

Many tasks in image processing are carried out by solving appropriate optimization problems. In the majority of these cases, the squared Euclidean distance is employed as a fidelity term. In light of our earlier comments, it is appealing to consider the SSIM as a fidelity term in such optimization problems, in an effort to enhance the visual quality of the solutions. That being said, the SSIM is not a convex function and therefore more difficult to handle mathematically, if not algorithmically.

Notwithstanding these obstacles, optimization problems that include the SSIM as a fitting term have already been

addressed. For instance, in [4] the authors find the best approximation coefficients in the SSIM sense when an orthogonal transformation is used (e.g., Discrete Cosine Transform (DCT), Fourier, etc.). Finding the best SSIM approximation coefficients is equivalent to minimizing the function

$$T(\Phi(x), y) = 1 - \text{SSIM}(\Phi(x), y), \quad (1)$$

where $\Phi(\cdot)$ is an orthonormal matrix, and y the signal being approximated. The function $T(x, y)$, which will be used in this paper, may be considered as a measure of the visual dissimilarity between x and y .

Based on the the results given in [4], in [6] Rehman *et al.* introduce the SSIM version of the the image restoration problem proposed by Elad *et al.* in [9]. Furthermore, in [6] the authors also introduce a super-resolution algorithm – also based on the SSIM – to recover from a given low resolution image its high resolution version.

Another interesting application for reconstruction and denoising of images is introduced by Channappayya *et al.* in [5]. In this case, the authors define the statistical SSIM index (statSSIM), which is an extension of the SSIM for wide sense stationary random processes. The non-convex nature of the statSSIM is overcome by reformulating its maximization as a quasi-convex optimization problem, which is solved using the bisection method [10], [5]. Nevertheless, it is not mentioned that the SSIM – under certain conditions – is a quasi-convex function.

More imaging techniques based on the SSIM can also be found in [7], [11], [8]. In these works, optimization of rate distortion, video coding and image classification are explored using the SSIM as a measure of performance.

Given that maximizing $\text{SSIM}(x, y)$ is equivalent to minimizing $T(x, y)$ in Eq. (1), it can be shown that all the applications mentioned above can be stated as the following optimization problem,

$$\min_x \{T(\Phi(x), y) + \lambda h(x)\}, \quad (2)$$

where Φ is usually a linear transformation, $h(x)$ is a regularizing term, and λ its corresponding regularization parameter. The constrained version of this problem is given by

$$\begin{aligned} \min_x \quad & T(\Phi(x), y) \\ \text{subject to} \quad & h(x) \leq \lambda. \end{aligned} \quad (3)$$

The methods used for solving (2) and (3) may converge to the same solution; if this is the case, it may be said that both optimization problems are equivalent.

Moreover, since many SSIM-based imaging tasks can be carried out by solving problems (2) and (3), we consider these equations as the general framework of what we call SSIM-based optimization. For this reason, we think that it is a better approach to develop algorithms for solving (2) and (3) rather than developing methods for addressing particular applications – which is the tendency found in the literature.

In this paper, we focus our attention on the quasi-convex approach; that is, we show how problem (3) can be solved. To do so, we show under what conditions the SSIM is quasi-convex. In addition, we consider the cases where $h(x)$ is convex and not necessarily differentiable. Furthermore, we show that (3) can still be solved if $T(\Phi(x), y)$ is subjected to a set of convex constraints [10]. Applications such as Total Variation and ℓ_1 norm constrained optimization are discussed, as well as comparisons between the ℓ_2 and SSIM approaches.

2. Structural Similarity (SSIM)

2.1 Definition

SSIM provides a measure of visual closeness between an image and a distorted or corrupted version of it. Since it is assumed that the distortionless image is always available, the SSIM is considered a *full-reference* measure of IQA [3].

The definition of SSIM is based on two assumptions: (1) images are highly structured – that is, pixels tend to be correlated, specially if they are spatially close – and (2), that the HVS is adapted to extract structural information. For these reasons, SSIM measures similarity by quantifying changes in perceived structural information. This measurement is done by comparing luminance, contrast and structure of the two images being compared. Changes in luminance are measured by quantifying relative changes in the means of the images; contrast comparison is carried out by measuring relative variance; finally, structure is compared simply by calculating the correlation coefficient between the two images. The SSIM is computed by multiplying these three factors together.

In what follows, and for the remainder of the paper, we let x and y denote two images or image patches with n components, i.e., $x, y \in \mathbb{R}^n$. The SSIM between x and y is then defined as

$$\text{SSIM}(x, y) = \left(\frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2} \right) \left(\frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \right) \left(\frac{\sigma_{xy}}{\sigma_x\sigma_y} \right), \quad (4)$$

which is equivalent to

$$\text{SSIM}(x, y) = \left(\frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2} \right) \left(\frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2} \right). \quad (5)$$

Here μ_x and μ_y denote the means of x and y , respectively, σ_{xy} denotes the cross-correlation between x and y and all other terms follow.

In order to avoid division by zero, positive constants C_1 and C_2 are added for purposes of stability, leading to the formula,

$$\text{SSIM}(x, y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right) \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right). \quad (6)$$

Since the statistics of images vary greatly spatially, $\text{SSIM}(x, y)$ is computed using a sliding window of 8×8 pixels. The final result, i.e., the so-called *SSIM index*, is basically an average of the individual SSIM measures.

2.2 Quasi-convexity

In this paper, we consider a special case of the SSIM defined in (6), in which both x and y are zero-mean vectors, i.e., $\mu_x = \mu_y = 0$. In this case, the luminance component, i.e., the first term in Eq. (6), is equal to one, so that the SSIM between x and y becomes

$$\text{SSIM}(x, y) = \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}. \quad (7)$$

Since $\mu_x = \mu_y = 0$, it follows that

$$\sigma_{xy} = \frac{1}{n-1} \sum_{i=1}^n x_i y_i \quad \text{and} \quad \sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2. \quad (8)$$

Using (8) and setting $C_2 = 0$, Eq. (7) reduces to

$$\text{SSIM}(x, y) = \frac{2x^T y}{\|x\|^2 + \|y\|^2}. \quad (9)$$

One might argue that removing C_2 is not good for the sake of stability. However, we shall assume that the observation y is always a non-zero vector. This ensures differentiability, stability and continuity of the SSIM for any $x \in \mathbb{R}^n$. Also, since the luminance component is not taken into account, no information is known about the non-zero mean optimal solution, which we shall designate as x^* . Nevertheless, in some applications, e.g. denoising of a signal contaminated with zero mean additive noise, the mean of y and x^* coincide, so that the non-zero optimal x^* can be recovered.

In order to show under what conditions $\text{SSIM}(x, y)$ is quasi-convex, we use the definition of quasi-convexity for a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$: f is quasi-convex if its domain and all its sub-level sets $S_\alpha = \{x \in \text{dom } f \mid f(x) \leq \alpha\}$, for $\alpha \in \mathbb{R}$, are convex [10]. The domain of $\text{SSIM}(x, y)$ is \mathbb{R}^n , which is convex. As for its sub-level sets, we have that

$$\begin{aligned} \text{SSIM}(x, y) = \frac{2x^T y}{\|x\|^2 + \|y\|^2} &\leq \alpha \\ -\alpha\|x\|^2 + 2x^T y - \alpha\|y\|^2 &\leq 0, \end{aligned} \quad (10)$$

where the latter expression describes a convex set as long as $\alpha \leq 0$. This implies that quasi-convexity exists if $x^T y \leq 0$.

Similarly, it can be shown that $\text{SSIM}(x, y)$ is quasi-concave if $x^T y \geq 0$.

Furthermore, due to the quasi-convexity and quasi-concavity of $\text{SSIM}(x, y)$, it follows that $T(x, y)$ in Eq. (1) is (i) quasi-convex if $x^T y \geq 0$ and (ii) quasi-concave if x and y are negatively correlated. In fact, from Eqs. (1) and (9), $T(x, y)$ may be expressed as follows,

$$T(x, y) = \frac{\|x - y\|^2}{\|x\|^2 + \|y\|^2}. \quad (11)$$

The range of this expression is the interval $[0, 2]$: $T(x, y) = 0$ when $x = y$ and $T(x, y) = 2$ when $x = -y$.

As mentioned before, since $\text{SSIM}(x, y)$ is a measure of similarity, $T(x, y)$ can be considered a measure of dissimilarity between x and y . In fact, $T(x, y)$ in Eq. (11) is an example of a (squared) normalized metric [12]. Since in the majority of optimization problems one minimizes a distance or an error subject to a set of constraints, we will define the SSIM-based optimization problems using the dissimilarity measure $T(x, y)$.

3. Optimizing the SSIM

We shall define a constrained SSIM-based optimization problem as follows,

$$\begin{aligned} \min_x \quad & T(\Phi(x), y) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b, \end{aligned} \quad (12)$$

where $\Phi(\cdot)$ is some linear transformation, $Ax = b$ is an equality constraint, and the $h_i(x)$ are a set of convex inequality constraints.

Assuming that the optimal zero-mean solution x^* is in the region where $T(\Phi(x), y)$ is quasi-convex, i.e., $(\Phi(x^*))^T y \geq 0$, the problem in (12) can be solved by solving a sequence of feasibility problems. For this, we require a family of convex inequalities that represent the sub-level sets of $T(\Phi(x), y)$ and a convex feasibility problem that is to be solved at each step. The bisection method may be employed to determine the optimal value of (12) up to a certain accuracy [10].

The family of convex inequalities is a set of functions $\phi_\alpha(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$f(x) \leq \alpha \iff \phi_\alpha(x) \leq 0. \quad (13)$$

Also, for every x , $\phi_\beta(x) \leq \phi_\alpha(x)$, whenever $\alpha \leq \beta$. The following functions satisfy such conditions:

$$\phi_\alpha(x) = (1 - \alpha)\|\Phi(x) - y\|^2 - 2\alpha(\Phi(x))^T y. \quad (14)$$

The feasibility problems then assume the form

$$\begin{aligned} \text{Find} \quad & x \\ \text{subject to} \quad & \phi_\alpha(x) \leq 0 \\ & h_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b. \end{aligned} \quad (15)$$

If (15) is feasible, then $p^* \leq \alpha$, else, $p^* > \alpha$, where p^* is the optimal value of (12).

Using the fact that $0 \leq T(\Phi(x), y) \leq 2$, and defining $\mathbf{1}$ and $\mathbf{0}$ as vectors in \mathbb{R}^n whose entries are all equal to one and zero respectively, we propose the following algorithm for solving (12):

Bisection method for constrained SSIM-based optimization

```

initialize  $x = \mathbf{0}$ ,  $l = 0$ ,  $u = 2$ ,  $\epsilon > 0$ ;
data preprocessing  $\bar{y} = \frac{1}{n}\mathbf{1}^T y$ ,  $y = y - \bar{y}\mathbf{1}^T$ ;
while  $u - l > \epsilon$  do
   $\alpha := (l + u)/2$ ;
  Solve (15);
  if (15) is feasible,  $u := \alpha$ ;
  elseif  $\alpha = 1$ , (12) can not be solved, break;
  else  $l := \alpha$ ;
end
return  $x$ ,  $y = y + \bar{y}\mathbf{1}^T$ .

```

Notice that this method will find a solution as long as $(\Phi(x))^T y \geq 0$; i.e., if the condition holds, the algorithm converges to an optimal value p^* . This condition may seem restrictive, however, one is normally interested in solutions that are positively correlated to the given observation y .

It is worthwhile to mention that it is not always possible to recover the mean of the non-zero-mean optimal solution x^* . This is because the luminance component of the SSIM has not been taken into account. Nevertheless, in many circumstances (e.g., denoising of a signal corrupted by zero-mean additive white Gaussian noise), the mean of y and $\Phi(x^*)$ coincide. In this case, $x^* = x^* + \hat{x}$, where x^* is the zero-mean optimal solution and \hat{x} is a vector such that $D\hat{x} = \bar{y}\mathbf{1}$. If $\Phi(\cdot)$ is any $m \times n$ matrix D , it can be seen that \hat{x} is given by:

$$\hat{x} = \bar{y}(D^T D)^{-1} D^T \mathbf{1}, \quad (16)$$

provided that the inverse of $D^T D$ exists.

4. Applications

Clearly, different sets of constraints lead to different SSIM-based optimization problems. For instance, by disregarding the equality constraint in (12) and defining $h(x) = \|Ax\|_2^2 - \lambda$, where A is either a real or complex $p \times n$ matrix, the result is an SSIM version of an optimization problem with a Tikhonov constraint. Nevertheless, we focus our attention on some very interesting applications that arise when the ℓ_1 norm is used: (1) SSIM with ℓ_1 constraint (2) total variation (TV) (3) deblurring and (4) zooming.

4.1 ℓ_1 -constrained SSIM-based optimization

Making the substitution $\Phi(x) = Ax$, where A is a $m \times n$ matrix and $x \in \mathbb{R}^n$ (although x may be complex), and by using the convex constraint $h(x) = \|x\|_1 - \lambda$, we obtain the following SSIM based optimization problem:

$$\begin{aligned} \min_x \quad & T(Ax, y) \\ \text{subject to} \quad & \|x\|_1 \leq \lambda. \end{aligned} \quad (17)$$

This particular problem is appealing because it combines the concepts of similarity and sparseness. As with the classical LASSO method [13], [14], the solution of (17) is also sparse. To the best of our knowledge, this is the first reported optimization problem where the SSIM is optimized having the ℓ_1 norm as a constraint.

4.2 SSIM and Total Variation

By employing the constraint $h(x) = \|Dx\|_1 - \lambda$, where D is a difference matrix and $\Phi(x) = x$, we can define informally a SSIM-TV-denoising method for one dimensional discrete signals. Given a noisy signal y , its denoised version is the solution of the problem,

$$\begin{aligned} \min_x \quad & T(x, y) \\ \text{subject to} \quad & \|Dx\|_1 \leq \lambda. \end{aligned} \quad (18)$$

Notice that instead of minimizing the TV norm, we employ it as a constraint. This approach is not new – it can also be found in [15], [16]

Moreover, images can also be denoised by minimizing the dissimilarity measure $T(x, y)$ subject to the following convex constraint:

$$h(x) = \|D_x(x)\|_1 + \|D_y(x)\|_1 - \lambda, \quad (19)$$

where the difference matrices $D_x(\cdot)$ and $D_y(\cdot)$ are linear operators used to compute the discrete spatial derivatives. Notice that the anisotropic TV norm is being used in this case. As far as we are concerned, the work reported in [17] and this application are the unique approaches in the literature that combine TV and the SSIM.

4.3 Deblurring

The blurring of an image is usually modelled as the convolution of an undistorted image x and a blur kernel τ . Nevertheless, in practice, the blurred observation y may have been degraded by either additive noise or errors in the acquisition process. For this reason, the following model is used to represent the degradation process [18], [19]:

$$y = \tau * x + \eta, \quad (20)$$

where η is usually white Gaussian noise.

The problem of recovering x can be addressed by the proposed approach by using the convex constraint $h(x) = \|D_x(x)\|_1 + \|D_y(x)\|_1 - \lambda$, and by defining $\Phi(x) = Kx$,

where K is a linear operator that performs the blurring process. That is, the unblurred image x can be estimated by solving the following SSIM-based optimization problem:

$$\begin{aligned} \min_x \quad & T(Kx, y) \\ \text{subject to} \quad & \|D_x(x)\|_1 + \|D_y(x)\|_1 \leq \lambda. \end{aligned} \quad (21)$$

4.4 Zooming

In this case, given an image y , assumed to be of “lower resolution”, we desire to find an approximation x to a higher resolution version of y . This inverse problem can be solved in a manner very similar to the one described in the previous section; that is, by defining $\Phi(x) = Sx$, where S is a subsampling matrix, and using the same convex constraint that is employed for the deblurring application. We claim that a good estimate of the high resolution image x is the solution of the SSIM-based optimization problem given by

$$\begin{aligned} \min_x \quad & T(Sx, y) \\ \text{subject to} \quad & \|D_x(x)\|_1 + \|D_y(x)\|_1 \leq \lambda. \end{aligned} \quad (22)$$

Observe that, in general, the matrix $S^T S$ is not invertible, therefore, equation (16) can not be used to recover the optimal non-zero mean solution x^* . Nevertheless, the mean of the low-resolution observation y can be used as a good estimate of the mean of the high resolution image that is being sought.

The problem of zooming using the SSIM approach has also been addressed in [6], in which sparse representations of non-overlapping blocks of the image are used in the reconstruction process; however, the variational approach is not considered. Methods that employ the TV norm for estimating the high resolution image x can be found in [20], [19], nevertheless, the fitting term is the commonly used square Euclidean distance. Problem (22) can be considered as a method that combines the SSIM and the variational approach for addressing this inverse problem.

5. Experiments

In a series of numerical experiments, we have compared the performance of optimization methods employing (i) the usual squared Euclidean distance and (ii) Structural Similarity as fitting terms. For simplicity, we refer to these methods as (i) ℓ_2 -based and (ii) SSIM-based methods, respectively. This is done by comparing the structural similarities between an undistorted given image and both ℓ_2 and SSIM reconstructions. The structural similarities are calculated using the definition given by (9). By averaging the SSIM values of all non-overlapping 8×8 pixel blocks, the total SSIM for each recovered image is obtained. The reconstructions are obtained by solving either a SSIM-based or an ℓ_2 -based optimization problem over each pixel block. Finally, for each application, the corresponding constraint $h(x)$ being employed is the same for all non-overlapping blocks.

In all the applications that are presented below, the estimated mean from each block is removed prior to processing. Once the zero-mean optimal block x^* is obtained, the optimal non-zero-mean block x^* is recovered by means of Eq. (16), except in the case of zooming. In this case, the means of the high resolution blocks are approximated by the means of their corresponding low resolution counterparts. This is necessary since quasi-convexity of $T(\Phi(x), y)$ is guaranteed for zero-mean vectors. This approach is also applied in the classical ℓ_2 -based optimization method even though it is not required, for the sake of fair comparison of the two approaches.

For the ℓ_1 -constrained optimization problems of Section 4.1, both (17) and its ℓ_2 version are solved over each non-overlapping block. Here, $\Phi(x) = Dx$, where D is a $n \times n$ DCT matrix and $x \in \mathbb{R}^n$ is the set of DCT coefficients that is to be recovered. As expected, the fitting term of the ℓ_2 counterpart of (17) is substituted by $\|Dx - y\|_2^2$. The result is that for a given λ , a sparse approximation problem is solved at each block. In this experiment, a 96×96 sub-image of the test image *Mandrill* was used.

For the TV-based problems of Section 4.2, (18) is solved over each non-overlapping block. Since we are working with images, the constraint in (18) is replaced with the constraint

$$\|D_x(x)\|_1 + \|D_y(x)\|_1 \leq \lambda. \quad (23)$$

The fidelity term $\|x - y\|_2^2$ is employed in the ℓ_2 -based version of (18). In this experiment, a 96×96 noisy sub-image of the test image *Mandrill*, corrupted with additive zero-mean Gaussian noise (AWGN) ($\sigma = 1/32$) was employed.

For the deblurring problem of Section 4.3, a blurred and noisy 104×104 pixel subimage of the test image *Lena* was processed. The reconstructions were obtained by solving problem (21) and its ℓ_2 version over each non-overlapping block. The blurring kernel employed was a Gaussian with unit standard deviation. The blurred image was also contaminated with AGWN with $\sigma = \frac{1}{64}$.

Finally, with regard to the zooming problem of Section 4.4, the estimated high resolution images are obtained by solving both problem (22) and its ℓ_2 version over each pixel block. The fitting term employed in the ℓ_2 -based method was $\|Sx - y\|_2^2$.

Some results of sparse reconstruction, deblurring and zooming are presented in Figures 1, 2 and 3, respectively. In each Figure are shown the original (uncorrupted) image, its corrupted version and SSIM- and ℓ_2 -based reconstructions. For each set of experiments, the SSIM maps between reconstructions and the original image are presented. The brightness of regions in the SSIM maps indicates the degree of similarity between corresponding image blocks – the brighter a given point the greater the SSIM, hence visual similarity, at that location [3].

In Figure 1, where results of the sparse reconstruction problem are shown, the SSIM- and ℓ_2 -based reconstructions

are very similar. However, we notice that the SSIM reconstruction enhances the contrast at some locations (e.g., the wrinkles below the eye of *Mandrill*).

With regard to Figure 2, where the deblurring results are shown, we see that edges in the reconstructions tend to be sharper than those in the blurred and noisy image. However, the reconstruction of textures is not very good in both methods. We expect that improved results may be obtained by tuning the constraints of each pixel block for optimal performance.

With regard to Figure 3, where zooming results are shown, we observe that both SSIM- and ℓ_2 -based reconstructions are quite good. Some vertical and horizontal artifacts can be seen in both reconstructions, located mainly near edges. This may be due to the fact that anisotropic total variation was employed as a constraint.

A summary of quantitative results is presented in Table 1. The effectiveness of both SSIM- and ℓ_2 -based approaches was quantified using the mean squared error (MSE) and the SSIM defined in Eq. (9). The best results with respect to each measure of performance are denoted in bold. We observe that the effectiveness of both methods is almost the same, although the proposed approach performs better respect to the SSIM measure $\text{SSIM}(x, y) = 1 - T(x, y)$, as expected. We also observe that a low MSE does not necessarily imply a high visual similarity (measured in terms of SSIM) between the reconstructions and the original images, as is well known in the literature [3].

6. Concluding remarks

We conclude by reminding the reader that the primary purpose of this paper was to establish a general framework for constrained SSIM-based optimization based on the quasi-convexity properties of the SSIM – something that has not yet been done in the literature. The experimental results presented above are preliminary and can, in no way, be used to claim that SSIM-based optimization yields superior results in terms of visual quality. Further investigations will have to be performed in order to examine this conjecture. Our paper provides both the theoretical and computational backgrounds to continue such investigations.

	PROPOSED		ℓ_2	
	SSIM	MSE	SSIM	MSE
SPARSE RECONS.	0.6779	7.6729	0.6656	8.0950
TV DENOISING	0.8903	1.5601	0.8894	1.5549
DEBLURRING	0.6911	2.0389	0.6878	2.0197
ZOOMING	0.8167	2.0753	0.8142	2.1067

Table 1: Numerical results for the different approaches and applications. Numbers in bold identify the best results with respect to each measure of performance.

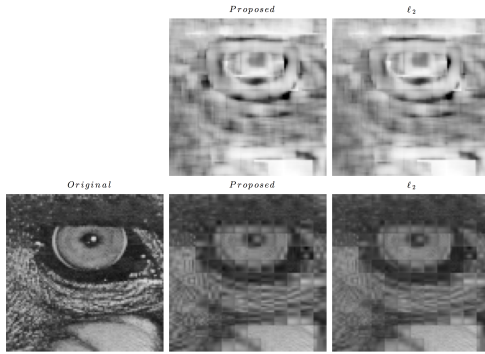


Fig. 1: Visual results for the sparse reconstructions. In this case, for each pixel block, the maximum allowed value for the ℓ_1 norm of the coefficients that are to be recovered is 1. In the top row, SSIM maps are shown. Original and recovered images can be seen in the bottom row.

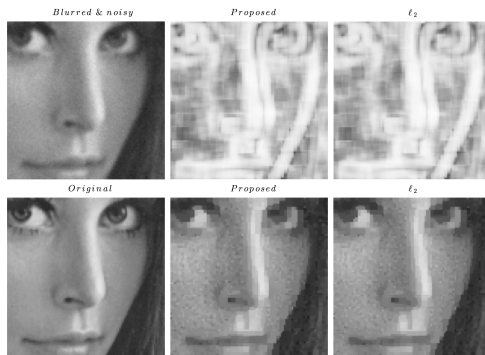


Fig. 2: Visual results for the deblurring application. In the top row, the blurred and noisy image along with the SSIM maps are presented. As above, the recovered and original images are seen in the bottom row.



Fig. 3: Visual results of the zooming experiments. The low resolution image and the SSIM maps can be seen in the top row. Original image along with the SSIM and ℓ_2 reconstructions are shown in the bottom row.

References

- [1] Z. Wang and A. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE*, vol. 9, no. 3, pp. 81–84, 2002.
- [2] G. Albuquerque, M. Eisemann, and M. A. Magnor, "Perception-based visual quality measures," in *IEEE VAST*. IEEE, 2011, pp. 13–20.
- [3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.
- [4] D. Brunet, E. R. Vrscay, and Z. Wang, "Structural similarity-based approximation of signals and images using orthogonal bases," in *ICIAR (1)*, ser. Lecture Notes in Computer Science, vol. 6111. Springer, 2010, pp. 11–22.
- [5] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. H. Jr., "Design of linear equalizers optimized for the structural similarity index," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 857–872, 2008.
- [6] A. Rehman, M. Rostami, Z. Wang, D. Brunet, and E. R. Vrscay, "Ssim-inspired image restoration using sparse representation," *EURASIP J. Adv. Sig. Proc.*, vol. 2012, p. 16, 2012.
- [7] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Ssim-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 22, no. 4, pp. 516–529, 2012.
- [8] A. Rehman, Y. Gao, J. Wang, and Z. Wang, "Image classification based on complex wavelet structural similarity," *Sig. Proc.: Image Comm.*, vol. 28, no. 8, pp. 984–992, 2013.
- [9] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [11] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on ssim-inspired divisive normalization," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1418–1429, 2013.
- [12] D. Brunet, "A study of the structural similarity image quality measure with applications to image processing," Ph.D. dissertation, University of Waterloo, 2012.
- [13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.
- [14] J. M. F. Bach, R. Jenatton and G. Obozinski, *Convex Optimization with Sparsity-Inducing Norms*. Optimization for Machine Learning, MIT Press, 2011.
- [15] P. L. Combettes and J.-C. Pesquet, "Image restoration subject to a total variation constraint," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1213–1222, 2004.
- [16] J. Fadili and G. Peyré, "Total variation projection with first order schemes," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 657–669, 2011.
- [17] Y. Shao, F. Sun, H. Li, and Y. Liu, "Structural similarity-optimal total variation algorithm for image denoising," in *Foundations and Practical Applications of Cognitive Systems and Information Processing*, ser. Proceedings of the First International Conference on Cognitive Systems and Information Processing, Beijing, China, Dec 2012 (CSIP2012), vol. 215. Springer, 2012, pp. 833–843.
- [18] P. Getreuer, "Total variation deconvolution using split bregman," *Image Processing On Line*, vol. 10, 2012.
- [19] B. Goldluecke, E. Strekalovskiy, and D. Cremers, "The natural vectorial total variation which arises from geometric measure theory," *SIAM J. Imaging Sciences*, vol. 5, no. 2, pp. 537–563, 2012.
- [20] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

Book spread correction using a time of flight imaging sensor

L. Galarza, Z. Wang, and M. Adjouadi

Center for Advanced Technology and Education, College of Engineering and Computing,
Florida International University, Miami, Florida, USA

Abstract - This paper proposes an approach that utilizes depth map information from a time of flight sensor in order to correct the warping in a book spread's image. This approach does not assume a uniform distribution of the pages of the book spread, but rather that there are varying levels of deformation due to changes in the curvature within each page. The corrections are achieved by modifying the lens equation to take into account different height points on the book spread. In addition, current resolution limitation from the time of flight device is overcome by scaling and matching the pixel depth data to an image from a camera with higher resolution. The Implementation results are shown in support of this approach.

Keywords: Time of flight; depth map; image correction; book reader

1 Introduction

The purpose of this suggested approach is the correction of a book spread. The approach relies on being able to obtain an accurate depth map of the book spread. Possible depth recovery devices include those that are based on stereo imaging, scatter light, laser scanning and time of flight. In this case, the time of flight device was chosen due to its higher processing speed and light weight while maintaining a good level of depth accuracy. These characteristics have led to a proliferation in the use of the time of flight devices in a number of applications [1]. A current drawback from this device is that it currently produces a low resolution, which can be overcome by pairing it with a higher resolution device, as has been proposed in other methods [2], [3]. The corrections will be made on the higher resolution image, where the warped effect is most noticeable.

There are a number of varying approaches to correct the warped effect on book spread images, which range from using geometrical constraints (such as cylindrical and linear characteristics), to optical flow and lens correction [4]-[8]. With this in mind, the approach taken will be one that is derived from lens correction. In the following sections, it will be shown how the height was incorporated as part of the modified lens equation to determine the unwarped pixel

locations. The validity of this approach will be made possible via an experimental setup, which should produce a flattening and expanding effect on the book spread image, to be demonstrated in the results.

2 Methods

2.1 Book spread correction

The location of an object relative to the camera can cause a warping effect. If the distance from the camera to the object remains fixed, then the factor affecting the display can be attributed to the height of the object, which in this setup would be a book spread. Therefore, the desired non-affine transformation to find the flatten location was derived based on the following conceptual design.

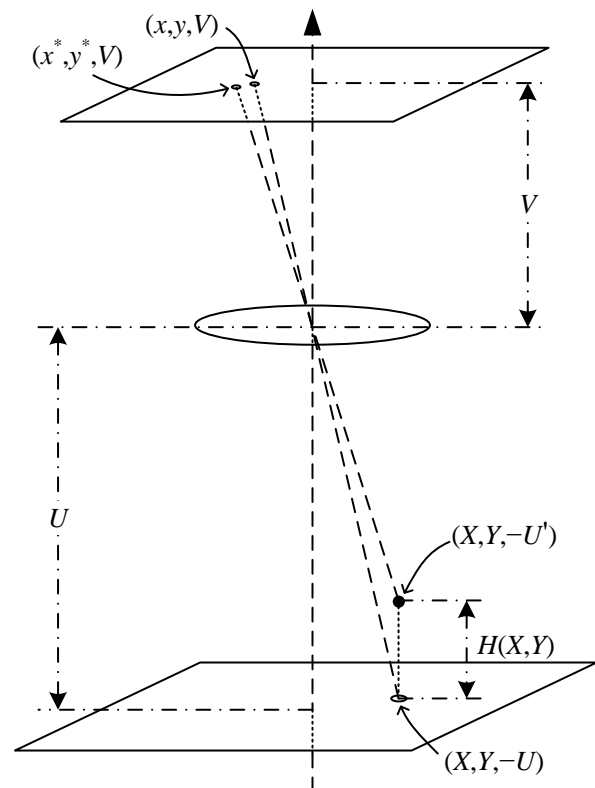


Fig. 1. Conceptual Design and Geometry of the Book reader

In Fig.1, V is defined as the distance of image sensor from the lens, U is the distance of the base of a platform from the lens, and point $(X,Y,-U')$ is a point along the book spread, with a height of $H(X,Y)$. When this location is captured by the camera, it yields the warped location (x^*,y^*,V) . However, the desired flattened location on the image sensor plane is at location (x,y,V) . Therefore, to realize the correction, a modified lens equation is used.

Using the properties of the lens and the geometrical layout illustrated in Fig. 1, the following relationship is established:

$$y/Y = x/X = V/U \quad (1)$$

Thus, by taking into account the height $H(X, Y)$, the warped coordinates could be derived as shown in equation (2):

$$\begin{aligned} x^*/X &= V/(U - H(X,Y)) \text{ , and} \\ y^*/Y &= V/(U - H(X,Y)) \end{aligned} \quad (2)$$

However, the location (X, Y) on the book spread is unknown and what is desired is the corrected location (x, y) in the captured image. Therefore, the height in terms of x and y can be expressed as

$$H(X,Y) = H(x \times (U/V), y \times (U/V)) = h(x,y) \quad (3)$$

Where $h(x,y)$ is the height to be recovered from the unwarped image. What is actually known is the warped location (x^*, y^*) . In order to obtain $h(x,y)$, the changes along x^* and y^* are considered minimal, allowing $h(x,y)$ to be expressed as

$$h(x,y) = h(x^* + \Delta x^*, y^* + \Delta y^*) \approx h(x^*, y^*) \quad (4)$$

Therefore, the corrected location (x,y) on the captured image plane can be expressed as

$$\begin{aligned} x &= x^* \times (U - h(x^*, y^*)) / U, \text{ and} \\ y &= y^* \times (U - h(x^*, y^*)) / U \end{aligned} \quad (5)$$

Applying this transformation along all points in an image will allow the image to be corrected and the warping effect on the book spread to be attenuated to a great extent as the results would confirm.

2.2 Experimental setup

The approach considered in this study heavily relies on the ability to obtain accurate height information. Therefore, the Argos3D-P100 [9] was selected as a viable alternative. This time of flight device is small in size, lightweight, and can provide gray scale image information while maintaining a good depth recovery performance. Ideally, a single device that

can obtain the pixels' amplitude and depth information would be desired. However, the native resolution of the Argos3D-P100 is only 160x120 pixels. The current deficiency of the time of flight device is overcome when the device is paired with a higher resolution camera, which in this case is the Canon G6. The placement of the time of flight device and higher resolution camera relative to the book spread can be observed in Fig. 2.

The time of flight device is positioned next to the higher resolution camera. This positioning will allow for an unobstructed view over the testing area. The book spread will then be captured by both devices to be appropriately matched later.

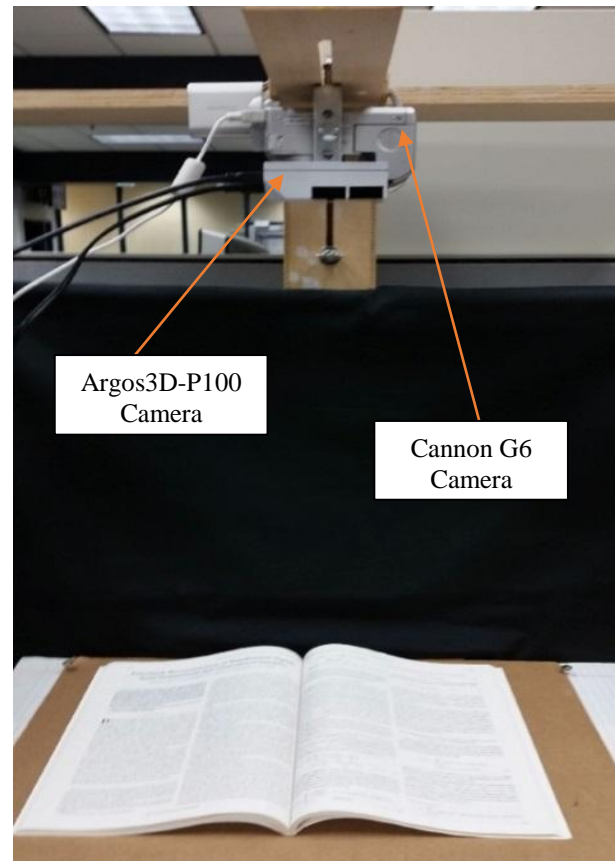


Fig. 2. Experimental Setup

2.3 Implementation

The first step to realize the implementation was to ensure the accuracy of the depth map via a calibration process akin to [10]. Specifically, for this application, the distance from the sensor to the surface of the book holder platform was measured and then compared to the depth data obtained from the calibrated Argos3D-P100. Also, as it has been demonstrated [11], to further ensure the depth resolution, the native bilateral filter was used. The approach will focus on the book spread area. Furthermore, to reduce inaccuracies due to depth sampled variances, the average of 50 depth maps samples were used.

In the next step, the pixel values are gathered from both devices along with depth information. The height of the book is then calculated by taking the averaged depth map from the book holder platform and subtracting it from the averaged depth map of the book spread. This height can be used directly so that it reflects the raw height as obtained from the Argos3D-P100 camera:

$$h(x^*, y^*) = \text{Raw Height} \tag{6}$$

However, another approach would be to use polynomial approximations of order n with coefficients p for each of the row of the height points as shown in equation (7):

$$h(x^*, y^*) = \begin{cases} p_{11}x^{*n} + p_{12}x^{*(n-1)} + \dots + p_{1n}x^* + p_{1n+1} & , y^* = 1 \\ \vdots & \vdots \\ p_{m1}x^{*n} + p_{m2}x^{*(n-1)} + \dots + p_{mn}x^* + p_{mn+1} & , y^* = m \end{cases} \tag{7}$$

The lower resolution book spread image is then matched to the higher resolution image. The scale factors (one factor for the rows and another factor for the columns) from the match are then determined. The height map, $h(x^*, y^*)$, (which can be obtained from (6) or (7)) is then linearly interpolated using the scale factors to match each of the corresponding pixels of the higher resolution image.

Lastly, the newly interpolated high-resolution height and amplitude pixel data is used in conjunction with equations (4) and (5) to make the corrections. However, in order to use these equations, the value of U needs to be specified, which can be measured or calculated experimentally. In this implementation, U was calculated by placing a single flat page with label points on the book holder platform and just changing the height to a known elevation. Thus, for this setup, U was determined to be 58.05 cm; other specific parameters from the higher resolution device are not required to perform this correction.

3 Results

As described in the implementation the Argos3D-P100 was calibrated and adjusted so that the book spread image obtained as seen in Fig. 1 where the resolution of the book spread image is 65x41.

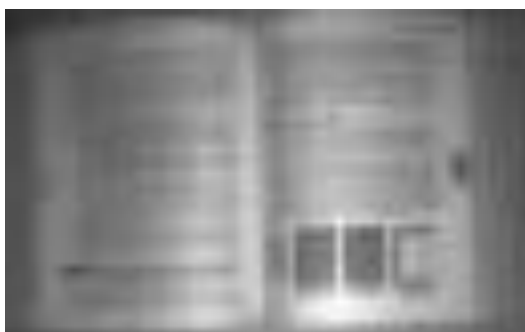


Fig. 3. Low resolution book spread

Combining the pixel values with the height values determined in equation (6) yields the results shown in Fig. 4.

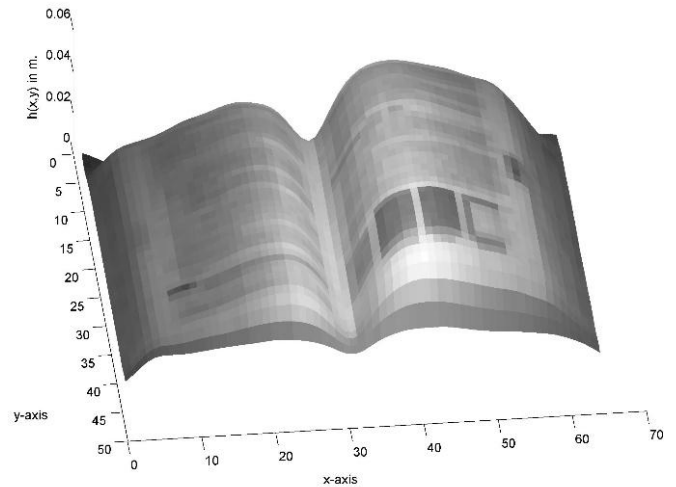


Fig. 4. Low resolution raw height map and pixel values

However, instead of using raw height values, a polynomial approximation (7) could instead be used as an alternative in order to provide continuous depth maps. A closer fitting can be achieved by increasing the order of the polynomial but will cause a reduction in the smoothness of the fitting if the order is too high. In this case, a 25th order polynomial was utilized and when compared with the raw height points the following was observed

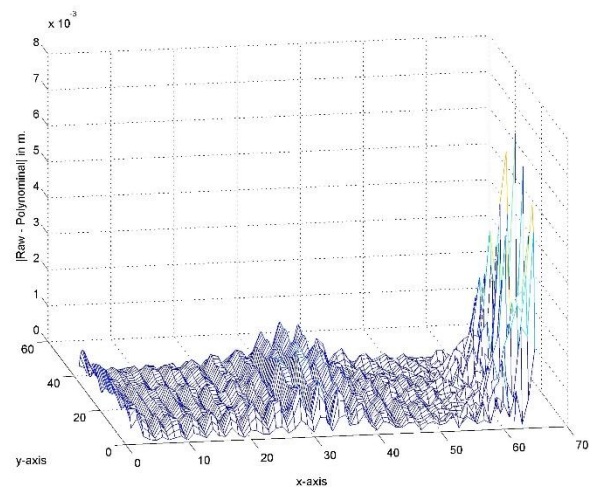


Fig. 5. Low resolution raw height map vs polynomial fitting

Fig. 5 shows that the most noticeable differences occur on the spine of the book spread and the edges. Using the polynomial height fitting on the low resolution image spread, as shown in Fig. 3, yields the results as shown in Fig. 6.

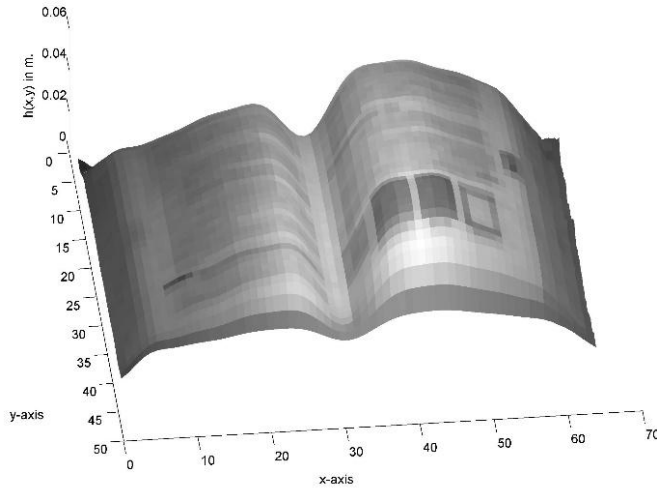


Fig. 6. Low resolution height polynomial fitting and pixel values

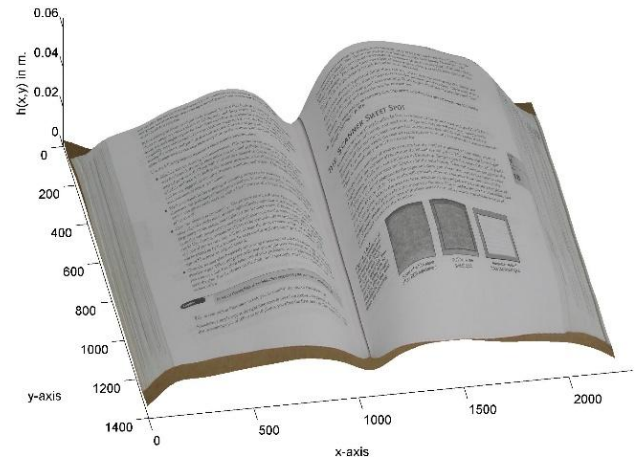


Fig. 8. Higher resolution raw height map and pixel values

As previously stated, the correction could be made ideally at this point. However, due to the low resolution of pixel values the text is illegible. Clear text recognition would therefore require a higher resolution image which necessitated the use of the Cannon G6 camera. The image resolution of the book spread in this case becomes 2226 x 1374, which is evidently more suited for OCR applications as is the case of this study.

The high resolution image in Fig. 7, as taken by the Cannon G6 camera, was thus matched to the lower resolution image shown in Fig. 3. Then the higher resolution height map (raw and polynomial fitted respectively) was calculated and combined with the higher resolution image to yield the results shown in Fig. 8 using raw data and Fig. 9 using polynomial curve fitting

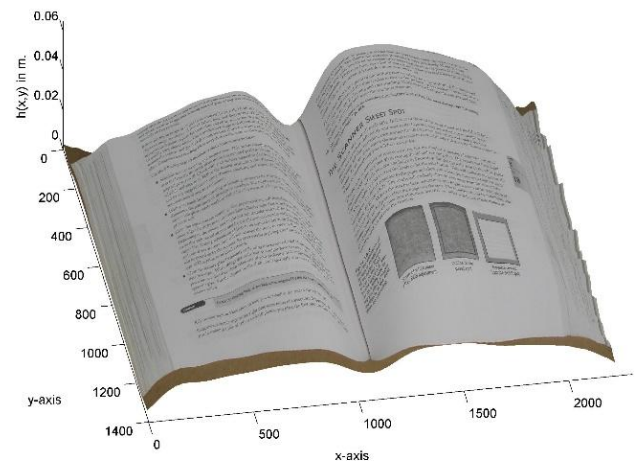


Fig. 9. Higher resolution height polynomial fitting and pixel values

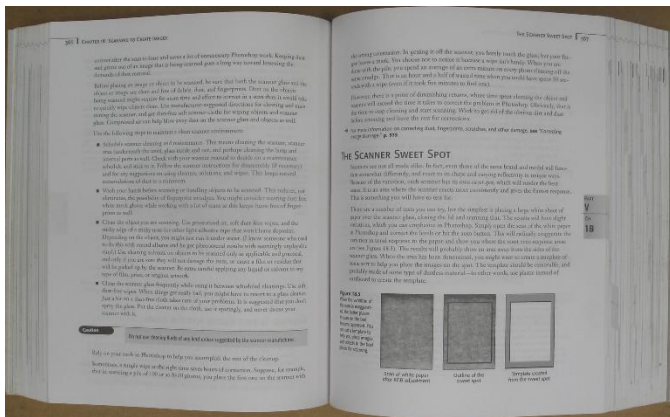
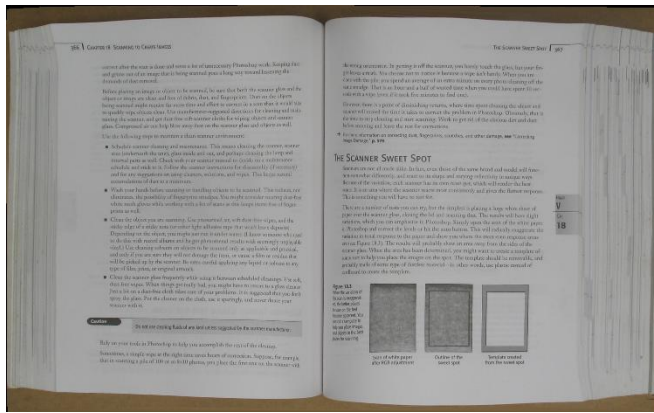


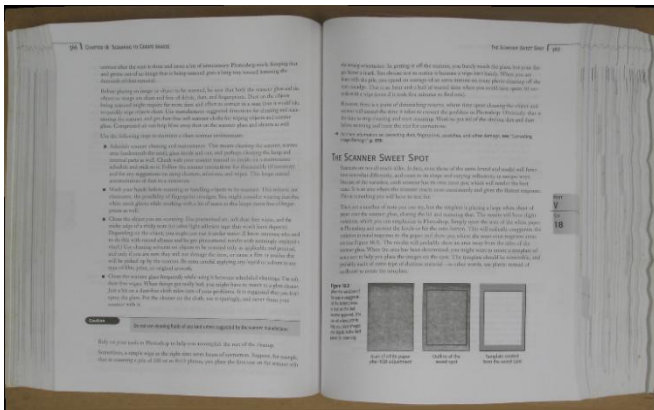
Fig. 7. Higher resolution image (warped)

At this stage, the corrections can be made. However, the resulting new locations were non-integer locations as opposed to the required integer positions for discrete images. Thus, a linear interpolation was used to determine the new pixel values at their corresponding integer locations. In Fig. 10, the book spread correction using the raw height values is shown in (a), while the height polynomial fitting correction is shown in (b)

Upon closer inspection of superior left corner of the book spread in each of the images on Fig. 7 and Fig. 10 respectively, it is possible to see how the warping effect is ameliorated and text is straightened as can be seen in Fig. 11.



(a)



(b)

Fig. 10. Corrections via (a) raw height & (b) polynomial fitting

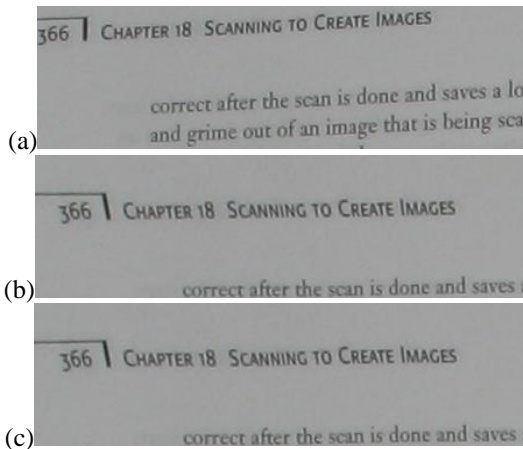


Fig. 11. Zoom location (a) warp, (b) correct with raw, and (c) corrected with polynomial fitting

Even under poor lighting conditions and with a thinner glossy book spread, the Argos3D-P100 was able to retrieve an adequate depth profile. The combined low resolution (47x62) raw height profile and pixel values can be seen in Fig.12.

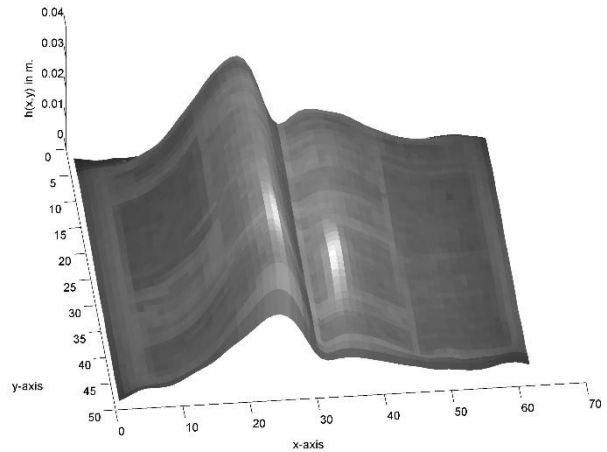


Fig. 12. Low resolution raw height map and pixel values

The bright areas as shown in Fig. 12 is where the glossiness had the most impact on the sensor. A comparison of the polynomial fitting versus the raw height values in this case can be seen in Fig.13.

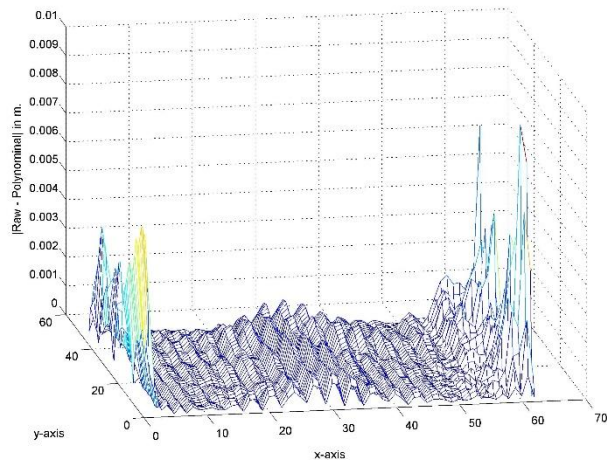


Fig. 13. Low resolution raw height map vs polynomial fitting

As was previously observed in Fig.5, discrepancies between the raw height and polynomial fitting appear to be on the edges and near the spine of the book spread. The combination of the polynomial fitting on the low resolution image can then be gauged as shown in Fig. 14 where the discrepancy of the edges is more pronounced. However, the target for correction is the higher resolution (2113x1603) image. Fig.15 shows the high resolution image, in which the lighting is not adequate and some of the text is affected because of it.

The next step was to interpolate the low resolution depth map and applying it to the higher resolution image. The combination of the high resolution image with the raw height map can be seen in Fig. 16, while the one by applying polynomial fitting can be seen in Fig. 17. Minor discrepancies from the previous low resolution combinations

in Fig. 12 and Fig.14 respectively are more evident in the higher resolution counterparts.

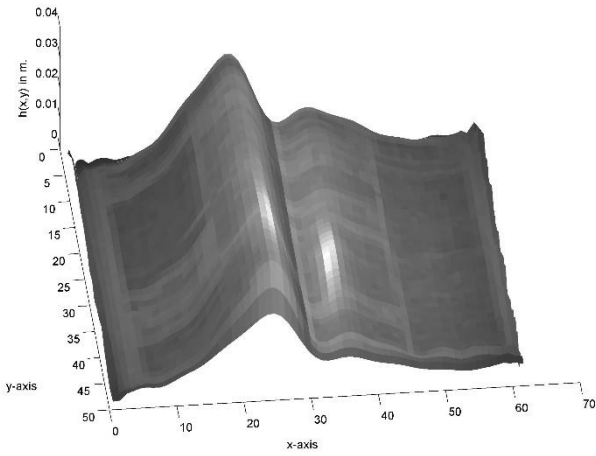


Fig. 14. Low resolution height polynomial fitting and pixel values

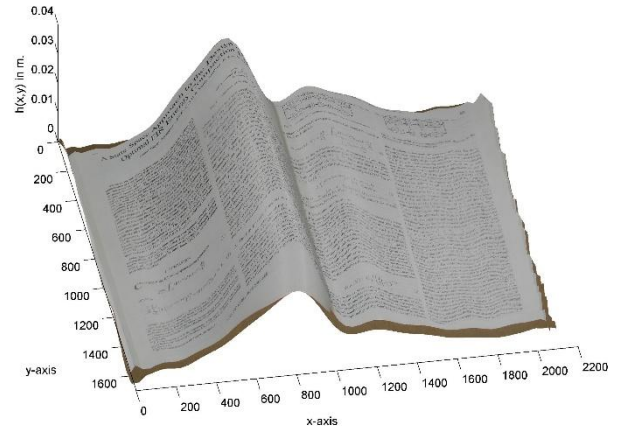


Fig. 17. Higher resolution height polynomial fitting and pixel values

Lastly, the correction was performed to reduce the warping effect on the book spread image and can be seen in Fig. 18.

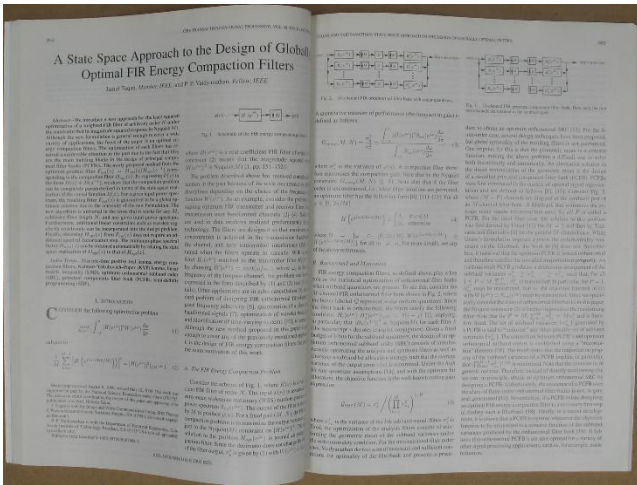
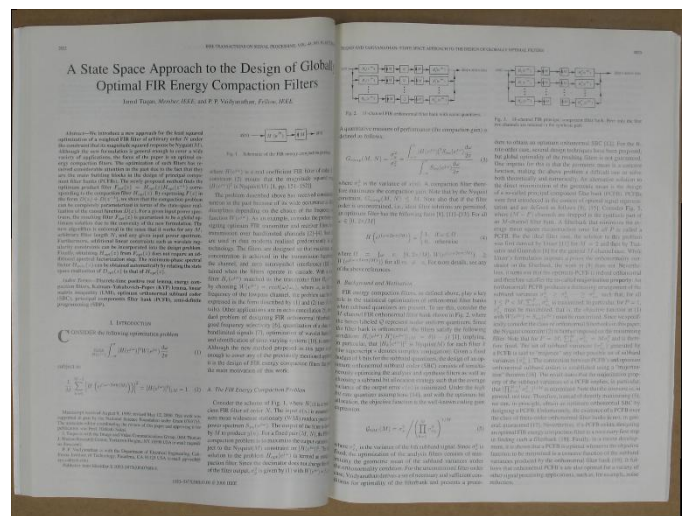


Fig. 15. Higher resolution image (warped) inadequate lighting



(a)

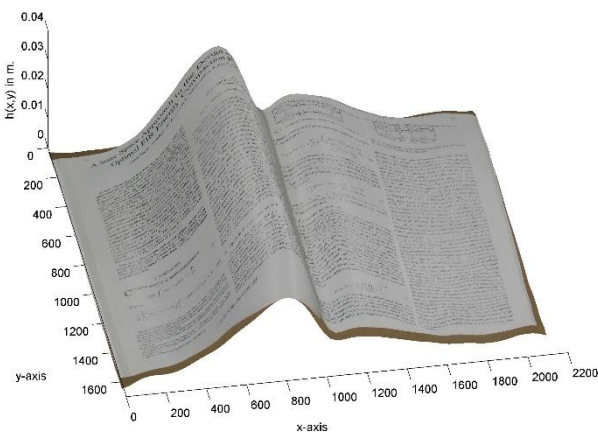
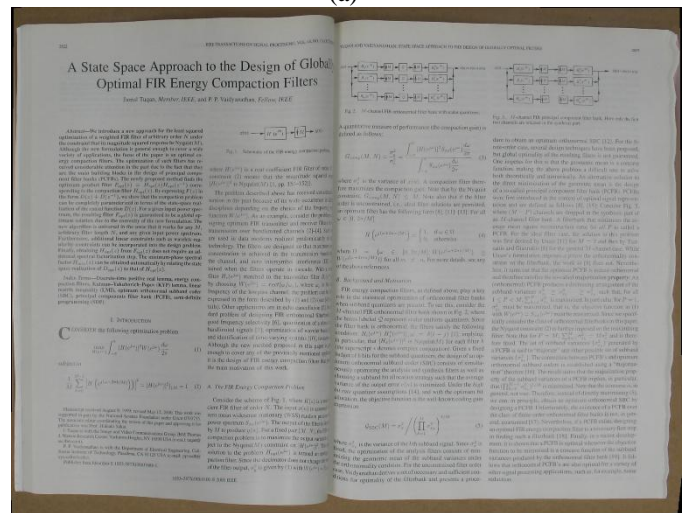


Fig. 16. Higher resolution raw height map and pixel values



(b)

Fig. 18. Images under poor lighting corrected with (a) raw height and (b) polynomial fitting

In order to better appreciate the effect of the correction, a small section of each case is shown in Fig 19.

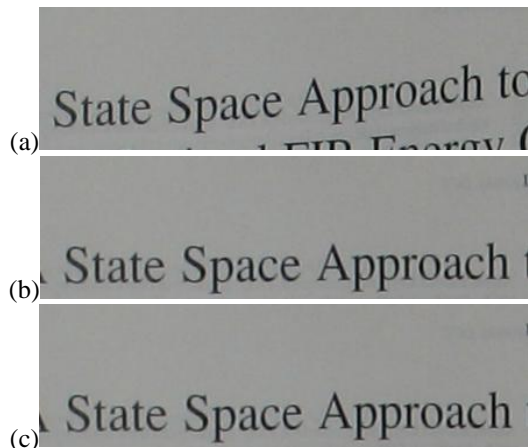


Fig. 19. Zoom location (a) warp, (b) correct with raw, and (c) corrected with polynomial fitting

4 Conclusions

The results showed that with the proposed mathematical derivations the image can be corrected and the warping effect is attenuated using both the raw depth data approach as well as the discrete polynomial curve fitting on the basis of the resolution of the 3D camera. The apparent curvature of the book is shown to have been flattened and extended correctly. However, the corrections are dependent on the height at each pixel point rather than assuming a uniform geometrical shape. Future work will include applying this correction to an OCR engine and testing its significance in terms of improvement in the text's recognition, as well as in ways to improve the results by extending the 1-D polynomial curve fittings into 2-D polynomial maps that will mimic the 3D depth maps of the raw data.

5 Acknowledgements

This research is supported through NSF grants CNS-0959985, CNS-1042341, HRD-0833093, IIP 1338922 and IIP-1230661. The support of the Ware Foundation is greatly appreciated.

6 References

- [1] S. Foix, G. Alenyà, and C. Torras, "Exploitation of time-of-flight (ToF) cameras," Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Tech. Rep. IRI-TR-10-07, 2010.
- [2] F. Garcia , B. Mirbach , B. Ottersten , F. Grandidier and A. Cuesta "Pixel weighted average strategy for depth sensor data fusion", Proc. IEEE 17th ICIP, pp. 2805-2808 2010
- [3] S. Schwarz, M. Sjöström, and R. Olsson, "A Weighted Optimization Approach to Time-of-Flight Sensor Fusion"

IEEE Transactions on Image Processing, VOL. 23, NO. 1, 2014

[4] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis. "A two-step dewarping of camera document images." In Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on, pp. 209-216. IEEE, 2008.

[5] J. Liang, D. DeMenthon, and D. Doermann. "Geometric rectification of camera-captured document images." Pattern Analysis and Machine Intelligence, IEEE Transactions on 30, no. 4 pp. 591-605. 2008

[6] L. Wang, and M. Adjouadi. "Automated Book Reader Design for Persons with Blindness." In Computers Helping People with Special Needs, pp. 318-325. Springer Berlin Heidelberg, 2008.

[7] M. Adjouadi, E. Ruiz, and L. Wang, "Automated Document Reader for Visually Impaired Persons", Springer's Lecture Notes on Computer Science (LNCS) series, K. Miesenberger et al. (Eds.): LNCS 4061, pp. 1094 – 1101, 2006.

[8] M. Cutter, and P. Chiu "Capture and dewarping of page spreads with a handheld compact 3D camera" In: Proceedings of DAS 2012, pp. 205–209, 2012

[9] K. Chelhwon, P. Chiu, and S. Chandra. "Dewarping Book Page Spreads Captured with a Mobile Phone Camera." In Camera-Based Document Analysis and Recognition, pp. 101-112. Springer International Publishing, 2014.

[10] https://support.bluetechnix.at/wiki/Argos%C2%AE3D_P100_Camera

[11] S. Fuchs and G. Hirzinger. "Extrinsic and depth calibration of ToF-cameras." IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008. , pp.1-6. , 2008.

[12] C. Kim , H. Yu and G. Yang "Depth super resolution using bilateral filter", Proc. 4th Int. CISP, pp.1067-1071 20112

Optical Pedometer: A new method for distance measuring using camera phones

Joakim Sjöberg¹, and Mia Persson¹

¹Department of Computer Science, Malmö University, 205 06 Malmö, Sweden

Abstract—Present day pedometer applications lack the ability to identify and measure each step for individual data with high precision. The rapid growth and evolution in the capacity of today's smartphones now present the opportunity to investigate new methods that gives the possibility to measure each individual step, providing data such as length, width and time by using an out of shelf smartphone. This paper introduces the Optical Pedometer (OP for short) that differs from traditional smartphone pedometer applications in the way that it is based on computer vision, providing new possibilities as to actually identify and measure steps with help of the phone's camera. Our proposed method was evaluated and compared to the existing methods within this field; accelerometer based pedometers and GPS applications. The results showed that the OP method presented more accurate measurements, thus proving it applicable for shorter measurements requiring a higher degree of accuracy.

Keywords: android, pedometer, computer vision, distance measuring

1. Introduction

Measuring traveled distances with your smartphone is a popular concept used in contemporary applications predominantly within areas concerning health, exercise and recreation.

To measure covered distances the system rely on either one or two of the phone's sensors: the Global Positioning System (GPS) sensor and the accelerometer. The GPS sensor in the phone sends information about the current position in form of coordinates. The length between the coordinates can then be calculated to give an account for the total distance traveled. The accelerometer, on the other hand, is often used to create a step counter or pedometer which identifies changes in the signal provided by the accelerometer to detect when a step has been taken [12]. The total number of steps is then multiplied with the users estimated average step length resulting in an assessment of the total distance walked.

However, none of these methods works well for measuring shorter distances to any high degree of accuracy. For example the GPS approach is used for most exercise applications and works well for longer distances as its accuracy improves the further the distance you travel. However, according to the American Department of Defense [2] the public civilian

GPS type called Standard Positioning Service (SPS) hold a horizontal accuracy of 7.8 meter (95 percentage Confidence interval) which makes it inapplicable for measurements of shorter distances with a high degree of accuracy. Another problem with the GPS method for measuring distances is the signal which easily gets obstructed by nearby buildings and trees which subsequently inhibits its indoor use.

With the pedometer method on the other hand distances are measured based on an average step length approach. The accuracy of the measurement will, in similarity to the GPS approach, increase as the distance covered increases. The reason behind this comes from simple statistics; as a sample size increases the measurement becomes more reliable. Besides missing the actual length of the steps it also has problems with false positives meaning that other movements of the phone could be identified as steps leading to incorrect results.

Thanks to the rapid growth and continued improvement of capacity in today's smartphones, according to Moore's law [8], new possibilities have arrived allowing more accurate ways of measuring distances, even as short as 10 centimeters. To overcome the inaccuracy of earlier methods for short distance measuring we propose a new competitive method based on augmented reality through computer vision and basic math to develop an Optical Pedometer (OP).

The introduction of computer vision as the foundation in a pedometer application gives an advantage to the aforementioned methods by actually identifying the person's feet and track their movement in comparison to each other and the distance between them. Algorithms and basic mathematics can then be applied to get data such as: number of steps taken, the average step length, the step median, and the total distance traveled. To the best of our knowledge the deployment of a smartphone computer vision based method for counting and measuring steps has not been investigated in previous research. The technique is interesting compared to traditional techniques because it provides accurate measurements even in distances below 10 meters and is nowadays available in out of shelf mobile phones.

To evaluate the OP we conduct a comparative study based on the following main aspects: the number of steps identified and the total distance traveled. These features will be compared to current pedometer and GPS methods for measuring distances. New data is also presented attained with the OP such as average step length, width and time.

2. Theoretical background

2.1 GPS tracking applications

Applications that rely on the GPS to measure traveled distance, do this by using the GPS sensor in the phone. The sensor can receive signals from the GPS network containing 24 satellites, the estimated location is then derived from a technique called trilateration, using three or more satellites which gives the estimate position of the receiver (our phones GPS sensor) in latitude and longitude coordinates [1]. The application then stores the user's different coordinates and can calculate the difference between them or plotting them to a map to show the traveled distance.

Limitations to this approach involve its short distance uses, since its accuracy improves with increased distances. This is because of the horizontal accuracy of the Standard Positioning System (SPS) used for smartphones which is, 7.8 meter with a 95 percentage Confidence interval according to the American Department of Defense [2]. This means that if you travel a distance of 10 meters, chances are that the results will have a high degree of error percentage to the actual distance traveled, while if you travel 1 kilometer the error percentage will be lower and therefore yield more accurate results. Another problem with the GPS method for measuring distances is the signal which easily gets obstructed by nearby buildings and trees which subsequently inhibits its indoor use [2].

2.2 Accelerometer based applications

The accelerometer sensor in smartphones is used to identify differences in acceleration. The signal is then analyzed in order to identify the Gait cycle, example of how this is done can be found in [7]. So in simple terms it works in similar manners to traditional independent pedometer devices in the meaning that it senses the user's movements and through this identifies when steps are taken.

The average step length of the user is then assessed by the user or through a calibration process, and used for the estimate assessment of the total distance walked. As Hoang et al [7] concludes, the method has limitations in terms of actually identifying steps mainly because the fact that each individuals walking pattern is not constant but varies throughout the session and also day by day and for person to person. It also has limitations in not being able to measure the actual length of the steps and thus any measurements are based on average step length or step time, leading to variations in accuracy from time to time. However, in similarity to the GPS tracking approach; accelerometer applications has a higher degree of accuracy with increased distance traveled simply because of the increased sample size which makes the average step more accurate and consequently the total distance traveled.

3. Our new proposed method, the Optical Pedometer

3.1 Project background

In this section the background information regarding the techniques and theories behind OP are presented and their relation to different key aspects and functions are explained.

The prototype is developed for the Android platform using Nvidia Tegra Android Development Pack 2.0 for Linux with the minor modification of an updated version of the included OpenCV library to version 2.4.8.

3.2 Android pixel coordinate system

To be able to position items on the screen, a coordinate system of x and y is used to map every pixel. It is also possible to retrieve the coordinates from objects shown on the screen. As Figure 1 shows, the android system has both x and y as 0 in the top left corner and the maximum values in the bottom right corner. This means that if we analyzed a frame from our camera with two equal objects e.g. our feet, we could safely assume that the object with the lowest x value on the picture also is furthest to the left and the one with lowest y value would be the one positioned in front of the other.



Fig. 1: Android coordinate system.

3.3 Computer vision

Computer vision is the field where software is used to calculate and extract the information from a picture or frame provided by a camera. Algorithms are then used to process the images after the desired objects or patterns. For instance in advanced robotics computer vision is well applied to navigate. The robot uses a camera as its eyes and software as its brain, trying to figure out what it sees and how to react to it [6].

OP is based on computer vision, meaning that it's the method through which it identifies feet and therefore also recognize steps.

3.4 Object identification

To make our prototype able to detect and measure steps, we first need it to be able to identify the user's feet. There are several techniques to identify objects through computer vision: you can analyze and search the retrieved images after a key item, a specific color or shape. We have chosen to identify the user's footwear through simple color detection which is based on the provided color blob example in the OpenCV android library.

To accomplish this we first need to determine what color the object, in our case the user's footwear is. This is accomplished by letting the user identify its feet on the phones display and then tapping on one of them. This will start our tracking color identification which calculates the average color from the surrounding pixels at the users tap position on the screen. From this the user's footwear is deduced to be equivalent to the two largest areas of linked pixels with the target color range. To identify which foot is which, we let the pixel area with the lowest x value coordinate be identified as our left foot and the second as our right foot.

There are many aspects of this method that bring limitations, e.g. the feet need to be of a unique color in respect to the surrounding in order to minimize the amount of false positives. However, the method is general and can be used for every kind of footwear, shape or size. This makes it an adequate choice to fulfill the evaluation of the OP method.

3.5 Object tracking

Object tracking can be described as looking at the position of the identified object on one frame and estimate the difference relative to its position on the previous frame. In our case we look at the relative position of each foot on each frame and the distance between them, to see how it compares to the same values on the previous frame.

It's through this process we are able to determine when a step is taken and the length of each individual step. To do this we need some rules to define when a step is taken (what is the start and the end of one step) and what value to assign for its length.

a) Step identification: To identify when a step is taken we first need to set a rule defining what constitutes a step. For this instance we choose to classify a step as the action in which one foot is in front of the other. When the left foot is in front of the right foot we call it a left step and the opposite condition for a right step. This gives us the possibility to separate between a left step and right step.

We can then use the relative position of each foot to identify when a step has occurred. E.g. our left step starts on the first frame where the left foot is in front of the right foot and it ends on the first frame where the right foot is in front of the left foot, hence initiating the right step. This

could be described with the simple pseudo code as shown in Algorithm 1.

b) Step data: Once step identification is achieved the next process is to determine the length of each step. On each frame, for every step, we calculate the length in pixels between the foremost part of the left foot to the foremost part of the right foot. We then let a variable keep track of our step length, if the number of pixels between the feet are greater than the variable, the variable will adopt the value of the current frame. This way we ensure that the step length is the maximum distance measured in pixels between both feet on all frames during the cycle of a step. This could be described with the simple pseudo code as shown in Algorithm 1.

The same method is also used to measure the step width, which is the longest distance measured between the feet along the x-axis during the cycle of one step.

To measure the step time, we store a Unix time stamp at the start of each step cycle. We then subtract the time stamp from a newly made one at the end of the step cycle, leaving us with the total time in millisecond of each step.

Algorithm 1 Identify step

```

while measuring do
  if left foot is in front of rightfoot then
    if !leftStep then
      saveStep
      leftStep = true
    else
      if steplength < newlength then
        steplength ← newlength
      end if
    end if
  else
    if leftStep then
      saveStep
      leftStep = false
    else
      if steplength < newlength then
        steplength ← newlength
      end if
    end if
  end if
end while

```

3.6 Reference size

We use basic proportionality in order to convert step length in pixels to a size on the metric scale in cm. Since the main object of our attention is the footwear attached to our feet, we have the possibility to get the actual length of the shoe by simply converting the shoe size into the metric scale. For instance the European shoe size is based on Paris points which are convertible to cm by using formula 1 [3].

$$(\text{Paris points}) = \frac{3}{2} \times \text{foot length (cm)} + 1.5 \text{ cm} \quad (1)$$

This way we can prompt the user for his or hers shoe size and get a relatively precise measurement that can be used

as a reference size for our conversion from pixels to cm. This is possible since we now know both the actual length in cm of the shoe and its length represented in pixels on our screen. If we take the pixel size of the shoe and divide it by its actual length in cm we get how many pixels per cm proportion. So when we then want to convert our step length and total distance from pixels to cm we now have a constant value which we can use for the conversion.

If the user wants even more precise measurement or is wearing socks or other types of footwear the size can also be entered directly in centimeters.

3.7 Calibration

There are several factors that can affect the outcome of one measurement to another for example the person's height or shoe size. Also, the height at which the phone is held over the floor will make the image representation of the feet larger or smaller depending on the distance and this will affect the reference size. Another factor we need to take into account is perspective distortion, which means that at wide angles and short distances the objects representation may look distorted or stretched which could affect the distance in pixels. Also the lenses between different cameras may have different specifications resulting in different object representation.

To make a general solution that applies to phones with different camera types, people with different height and feet size calibration is needed. The calibration is done by putting one foot forward reaching the middle of the screen and the other one left back at the end of the screen. When the button is pushed the feet's total pixel size are added up by taking its lowest pixel coordinate's y value minus its biggest and divided by two to identify the average pixel representation of one foot (M) with equation 2 where f_1 and f_2 denote the pixel representation of the feet.

$$M = \frac{f_1 + f_2}{2} \quad (2)$$

This will minimize the distortion error from the distance equation. Because the foot which is left behind will be distorted and therefore showing a bigger size compared to the one positioned at the screen's middle. It will also adjust the reference size to match the persons height and shoe size.

Even if the calibration helps minimizing and eliminating some error factors it will not have any effect on the error happening during the actual measurement. For instance the movement of the phone along the vertical plane while measuring.

4. Data selection and validation

In this section we will evaluate the OP capabilities to measure steps but also the individual methods within the OP will be evaluated. In section 4.4, 4.5 and 4.6 equation 3 for standard deviation is used and all the tests are done

from our largest dataset from the 100 meters distance test in section 4.3.

$$s_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (3)$$

4.1 Calibration

To evaluate if our calibration method is working, we let two persons measure one step of 40 cm against a folding ruler as a reference to the actual length. The two test subjects will be equipped with different phones and be of different length and feet size as shown in the list below. As there is a low error margin shown for both test persons we conclude that our calibration method is working satisfactorily for the purpose intended.

[Test 1]

Person 1: length 180 centimeter, feet size 26.5 centimeter
Phone: LG G2, 13Mega pixel camera, camera lens F2.4, screen resolution 1080x1920
Reference size: 15.76
Result: 39.46 cm
Error margin: 1.35 %

[Test 2]

Person 2: length 165 centimeter, feet size 24.5 centimeter
Phone: Samsung S3, 8Mega pixel camera, camera lens F2.6, screen resolution 1280x720
Reference size: 11.59
Result: 40.28 cm
Error margin: 0.7 %

4.2 Pedometer

To evaluate our OP's capability to identify steps we compare it to the accelerometer based pedometer application according to the 20 steps test [9]. Also, we will perform a 50 steps and 100 steps test on the same test credentials as the 20 steps test with a maximum of 5 % error rate as acceptance limit. As shown in Table 1, the accelerometer based pedometer failed two out of three tests. OP on the other hand has a 0.0% error percentage in all three tests.

Table 1: Step identification test.

Steps	Optical Pedometer		Pedometer	
	result	error (%)	result	error (%)
20	20	0.0	17	15
50	50	0.0	57	14
100	100	0.0	102	2

4.3 Total distance traveled

To evaluate the distance measuring capabilities of OP we performed a series of test with OP, GPS and the Pedometer application on a range from one centimeter to 100

meters. The error percentages from the true values were then calculated for each of the three methods. For tests done with the GPS method Runkeeper was used [5] and for the accelerometer based pedometer Runtastic pedometer was used [4].

For measurements below 10 meters a measuring tape is used and for all other a running track is used as a reference size.

As illustrated in Table 2 below, our findings showed that neither the Pedometer nor the GPS based applications were able to measure distances below 10 meters making OP the only application with this capability. However OP error percentages on distances below 10 centimeters are too high to be trust worthy. We also see that OP remains below a 6% error rate in the range 10 centimeters to 80 meters. Where the other methods has a lot of variety in their results. This is mainly because they are only able to give estimates in the 10 meters range results, mainly because they are only able to give estimates in the 10 meters range.

The possibility to use a similar rounded values approach for the OP method is of course possible. However, not enough tests have been done to be able to validate an approach of this sort and therefore we chose to present the result as they are.

However we should remember that the dataset used for this analysis has a total distance error of 9.1%. Since our dataset contained 149 steps and the actual distance traveled was 100 meters our average step should be around 67 centimeters. We have no way to determine if the error is equally spread among the steps or if there are few steps which contains all the error. However if we deduct 9.1% of our mean step we get a new mean value of 66.55 centimeters which is closer to the previously mentioned value. Despite this we find the OP method qualified for estimate a person's average value with a relatively low error percentage which we estimate to be below 10%.

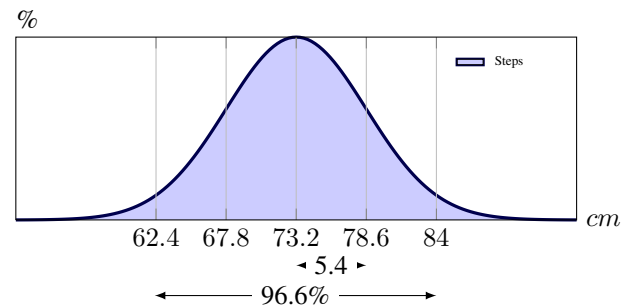


Fig. 2: Normal distribution step length test.

Table 2: Distances measured with different methods.

(m)	OP		GPS		Pedometer	
	result	error (%)	result	error (%)	result	error (%)
0.01	0.02	100	-	-	-	-
0.05	0.06	20	-	-	-	-
0.1	0.101	1	-	-	-	-
0.2	0.208	4	-	-	-	-
0.5	0.57	1.4	-	-	-	-
1	1.02	2	-	-	-	-
5	5.02	0.4	-	-	-	-
10	10.06	0.06	20	100	20	100
60	63.47	5.7	60	0.0	50	16.6
80	84.40	5.5	80	0.0	80	0.0
100	109.1	9.1	90	10	90	10

4.4 Average step length

To be able to evaluate the OP's capability to measure step length, standard deviation was calculated. Then both left and right steps were individually analyzed through the same method and test sample. No data was excluded from the population during this calculation and it was calculated as uncorrected sample standard deviation. The results are visualized in Gauss curves as shown in Figure 2 and 3 for all steps respectively left vs. right steps.

Our analysis shows that our median step has a length of 74,68 centimeters, and as shown in Figure 2 the mean step is 73.2 centimeters which compared to findings of previous research is indicated to be around 79 cm for males [11]. Which we take as a sign that the OP is well equipped for measuring average steps.

Figure 3 shows that the mean length for all steps are 73.2 centimeter, the diagram also illustrates that the left steps are slightly shorter, at 73.1 cm, compared to the slightly longer right steps, at 73.3 cm. Meaning that the person has a slightly longer right step than left step. However, if we look at the peak of the curves we see that the deviation is also slightly higher for the left steps with 6 centimeter respectively 4.7 centimeter for the right steps. After analyzing the results and dataset we can conclude that the OP is fully capable of analyzing an average step. Nonetheless the method requires active measuring from the test person, meaning it is not applicable for jogging or walking in high pace as it demands the person's attention.

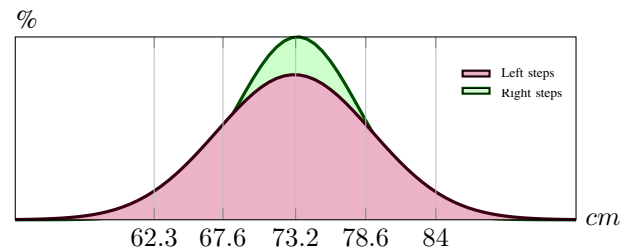


Fig. 3: Normal distribution step length left vs. right.

4.5 Average step time

To evaluate our step time we use the same approach as with our step length evaluation and furthermore base them on the same dataset. However the first step is removed from the population resulting in one sample less. This is because the first step is a part of the calibration process and therefore it's duration is longer than the average step which will affect the outcome of our calculations.

As seen in Figure 4 our mean step time is calculated to 703 ms, which if combined with our average step length could give us an average speed. The length 73.2 cm in 1.43 sec gives us an average speed of 3.75 km/h. This could be considered a slow speed for walking, which could be compared to that of a pedestrian walking which is found to be around 5.42 km/h for people below 65 years of age [10]. This is related to the current state of OP's development which demands a slower pace for more accurate measurements.

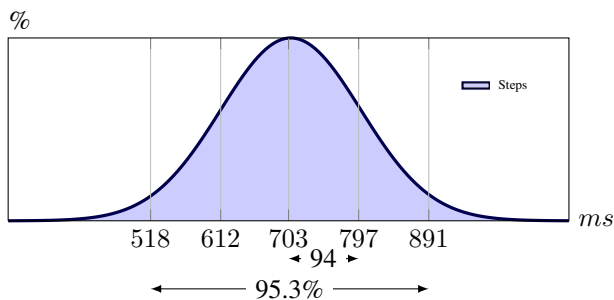


Fig. 4: Normal distribution step time.

When comparing the mean step time of left vs. right as shown in Figure 5 we see that the two curves have almost the same height, meaning that the standard deviation for both are similar. Which we found to be 104 milliseconds for left steps respectively 93 milliseconds for right steps. We can tell from the means that the left steps takes a slightly longer time compared to the right steps, with a mean value of 719 milliseconds respectively 699 milliseconds for the right steps. This in contrast to the previously shown difference in lengths between them could indicate that our test person were walking with a slight limp.

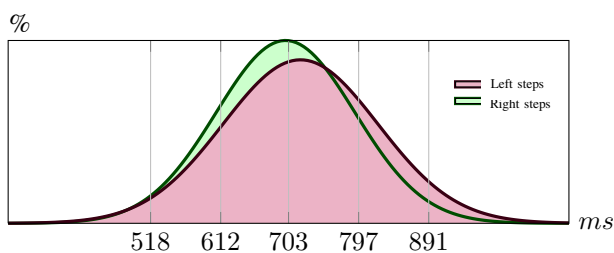


Fig. 5: Normal distribution step time left vs. right.

4.6 Step width

When analyzing our 100 meter dataset in regards to step width, we found that the mean step width was that of eight centimeter, and with a standard deviation of three centimeters. This gives us a variance of nine centimeter, and as shown in Figure 6 92.6 percentage of our steps where within two standard deviations.

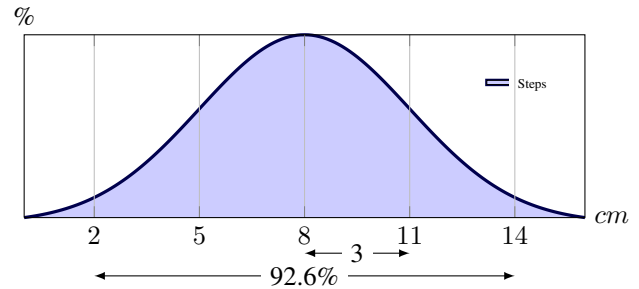


Fig. 6: Normal distribution step width.

5. Discussion and open problems

The Computer vision approach for developing a pedometer application has proven to be functional and possible to use with out of shelf smartphones. The method does provide unique possibilities to see and measure distances between feet, both in terms of length and width.

However in order to create a solution that is usable for the wider public, another method for identifying feet is necessary. The color identification approach is too sensitive to be practical in everyday situations because of the implication of the surroundings. For instance the color of the ground, shoe or pants and to some degree the lightning will affect the stability to correctly identify the feet.

Another downside for the computer vision approach is that it requires a lot from the phones resources, making the phone run a little slower than real-time pace. Therefore, the actual walking speed during the measuring process will have to be that of a slower pace. This allows the tracking process to view as many frames as possible during each step.

The basic proportionality of our reference size has been proven to be a valid method for conversion between pixels to the metric scale. However this approach works only because of the parallel and constant positioning of the camera over the feet. Any movement along the vertical plane or tilt of the phone would make the reference size misleading. A possible solution would be to perform an automatic calibration for each step or calculate an average reference size for each step based on its containing frames. Perhaps, also apply math with readings from the phones gyroscope to reduce some errors caused by tilting during measuring, this however is out of scope for this paper.

Our algorithm for step identification has proven to be very efficient and with a low to none error percentage. Regarding to step identification capabilities of pedometers the OP could

be considered as the one to beat with a 0 error percentage in our tests.

To summarize, the method works but is sensitive and puts restraints on the user, in terms that it requires an active measuring. Hence, to achieve accurate results the user must have the phone moving as little as possible in the vertical and horizontal plane while walking in a pace that allows the software in the phone to measure the steps accurately. However, this could be a sacrifice worth making if the user wants to measure short distances or get detailed information about left vs. right steps or step width not available by any other method.

References

- [1] R. Bajaj, S. L. Ranaweera, and D. P. Agrawal, "GPS Location Tracking Technology," *Computer.*, vol. 35, no. 4, pages. 92–94, Apr. 2002.
- [2] *Global Positioning System Standard Positioning Service Performance standard 4th Edition*, Department of Defense USA, 2008.
- [3] *Shoe sizes: Fundamentals of system size*, German Standard, DIN. 66074, 1975.
- [4] Google. (2014) Runtastic pedometer homepage on Google play. [Online]. Available:
<https://play.google.com/store/apps/details?id=com.runtastic.android.pedometer.lite>
- [5] Google. (2014) Runkeeper homepage on Google play. [Online]. Available:
<https://play.google.com/store/apps/details?id=com.fitnesskeeper.runkeeper.pro>
- [6] D. Hrach, and M. Brandner, "Intelligent Vision-Sensor for Robot-Sensing Applications," in *IEEE International Workshop on Robotic and Sensors Environments*, Ottawa, ON, Canada, pp 37–42, 2005.
- [7] T. M .Hoang, V. Q. Vo, T. D. Nguyen and C. Deokjai, "Gait Identification Using Accelerometer on Mobile Phone," in *International Conference on Control, Automation and Information Sciences IEEE*, vol. 24, no. 2, pp 344–348, Nov 2012.
- [8] C. A. Mack, "Fifty Years of Moore's Law," *IEEE Transactions on semiconductor manufacturing.*, vol. 24, no. 2, May 2011 pp 202–207.
- [9] C. Tudor-Locke, S. B. Sisson, S. M Lee, C. L Craig, R. C. Plotnikoff, and A. Bauman, "Evaluation of Quality of Commercial Pedometers," *Canadian Journal of Public Health*, vol. 97, pp. 10–15, Mar. 2006.
- [10] TranSafety. Inc. (1997). "Study Compares Older and Younger Pedestrian Walking Speeds." [Online]. Available:
<http://www.usroads.com/journals/p/rej/9710/re971001.htm>
- [11] D. Thompson, "Stride analysis (2002)." [Online] Available:
<http://moon.ouhsc.edu/dthomps/gait/knematics/stride.htm>.
- [12] W. Shyi-Shiou, W. Hsin-Yi, "The Design of an Intelligent Pedometer using Android," in *Second International Conference on Innovations in Bio-inspired Computing and Applications IBICA*, Dec .2011. pp 313–315.

Feature Enhancement of Robust Adaptive Target Detection with the Y-configured Multisensor Imaging Radar

Y. Shkvarko and V. Espadas

Department of Electrical Engineering, CINVESTAV Unidad Guadalajara
Av. Del Bosque 1145 Col. El Bajío, Zapopan, Jalisco, México
E-mail: shkvarko@gdl.cinvestav.mx

Abstract—A new Descriptive Experiment Design Regularization-based Robust Adaptive Beamforming (DEDR-RAB) approach is presented for high resolution array radar imaging of multiple targets with the Y-configured Multisensor Imaging Radar (MIR). Our approach is based on the advanced minimum risk inspired DEDR framework for enhanced radar imaging and optimization of the MIR resolution performances. We adopt the celebrated GeoSTAR sensor array configuration that provides a desirable low side lobes shape of the point spread function (PSF) attained employing the conventional matched spatial filtering (MSF) technique for radar image formation. The effectiveness (signal to interference plus noise ratio, SINR) of the new aggregated DEDR-RAB radar imaging method is corroborated via extended simulations of different DEDR-related imaging techniques. The results are indicative of the superior operational efficiency of high resolution localization of the multiple closely spaced targets with the new DEDR-RAB methodology.

Keywords: descriptive experiment design regularization, target detection, beamforming, multi-sensor imaging radar

1. Introduction

Beamforming is a pervading task in remote sensing (RS) imaging applications. The adaptive beamformers (ABF) select a weight vector as a function of the data to optimize the performance subject to various constraints, these ABF can have better resolution and much better interference rejection capability than the data-independent beamformers. However the ABF are much more sensitive to errors, such as steering vector errors caused by imprecise sensor calibration than the data-independent beamformers [1]. The latter has spurred development of various ABFs and devise RABs for enhancing the RS images, and many sophisticated techniques are now available (see, for example [1], [2] and the references therein). Crucial still unresolved ABFs issues relate to robust enhanced imaging in harsh operational scenarios characterized by possible imperfect array calibration, partial sensor failure and/or uncertain noise statistics [3].

We address a new DEDR-RAB approach for attaining virtual high-resolution performances of radar imaging with differently configured mm-band array radars. Our

new aggregated DEDR-RAB approach is a robust adaptive beamforming-oriented generalization of the conventional MSF method for radar image formation [4] based on the advanced descriptive experiment design regularization framework for radar imagery enhancement [5], [6]. At an initial stage, we optimize the sensor array configuration employing the celebrated GeoSTAR geometry [7] to attain the desired shape of the MSF system PSF, that is, we secure lowest possible side-lobes level balanced over the minimum effective width of the main PSF beam by optimizing the antenna inter-element spacing. At a reconstructive stage, the low resolution MSF image is next enhanced via performing the new aggregated DEDR-RAB post-processing aimed at attaining the overall super-high resolution remote sensing (RS) performances. The effectiveness of the new aggregated DEDR-RAB radar imaging method is corroborated via extended simulations of the multiple target localization experiment of different DEDR-related imaging techniques using the specialized elaborated software that we refer to as 'Virtual Remote Sensing Laboratory' (VRSL) [8]. The latter are indicative of the superior operational efficiency of the imaging radar system that employs the new DEDR-RAB method adapted to the DEDR-optimized GeoSTAR configuration over other tested competing techniques [1] - [9]. In this study, the robustness of the ABF is performed by aggregating it with the DEDR-based minimum variance distortionless response (MVDR) digital beamforming approach that exploits structural information on the desired image map sparsity over the RS scene [10]. The developed aggregated DEDR-RAB technique is implemented in an implicit iterative form to enhance the overall imaging and target localization performances.

2. MIR concept

The MIR-Y antenna array is shown in Fig. 1a and corresponding uv samples (Fig. 1b), in this case, u and v specify the normalized (so-called visibility domain) coordinate representation format, $u = x/\lambda_o$, and $v = y/\lambda_o$. This MIR array (GeoSTAR, Geo Synthesized Thinned Array Radiometer) is composed of 24 ($M = 24$) antenna elements as in [7], where it is addressed in as a concept to provide high resolution imaging of distributed RS scenes in

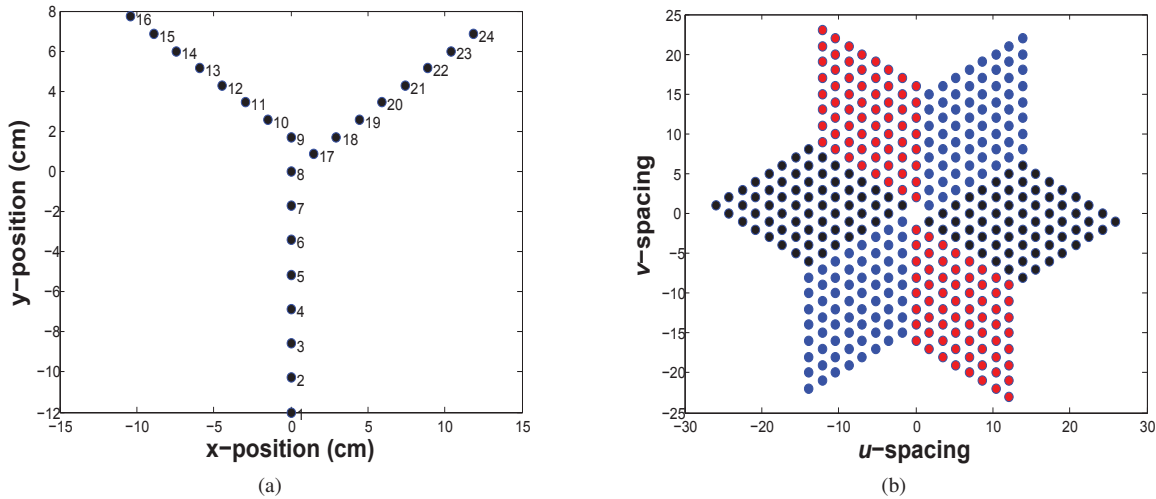


Fig. 1: (a) Antenna array layout with sensor numbering ($M = 24$) for Y-shaped GeoSTAR configuration, (b) corresponding uv samples for inter-element spacing $d_{A(1)} = 0.5\lambda_o$; carrier frequency $f_o = 24\text{GHz}$

microwave and mm wavebands. In this work, the particular system under consideration is operated at two separate yet concurrent frequencies of 24 GHz and 36 GHz with dual polarization (Vertical V and Horizontal H). At one instant, radio frequency (RF) pulses of a specified pulse width (PW) are transmitted concurrently at 24 and 36 GHz in either V polarization or H polarization. These pulses are “calibrated” to maintain coherency so that their amplitudes and phases are constant for different pulses. The transmitting antenna is switched between V and H polarizations; i.e., V and H transmitted pulses are delayed by a certain time. For each frequency (24 or 36 GHz), transmitted V polarized and H polarized RF pulses are separated by a half of the fixed pulse repetition time (PRT/2). The V polarized RF pulses and H polarized RF pulses are repeated after every PRT.

Each antenna element receives signals of V and H polarizations. It follows that, we can send V pulses and receive the same polarization mode (VV) or receive H polarization (VH); similarly, we have HH and HV modes. The operation range of the MIR system is in the order of 1 m to 100 m, with a range resolution cell of 0.3 m, so we have 165 range gates for processing. The sensors provide two measurements for each data snapshot, In-phase (I) and Quadrature (Q). The crucial issue relates to the formation of the empirical estimate (\mathbf{Y}_r) sensor data cross-correlation matrix (\mathbf{Y}_r) for each range gate $r = 1, \dots, R_r = 165$. To form the full rank cross-correlation matrix (\mathbf{Y}_r) we need to perform averaging over a great number J of independent recorded sensor array data realizations. These independent realizations are to be recorded using J transmitted pulses for each range gate $r = 1, \dots, R_r = 165$

To form the full-rank sensor data covariance matrix (\mathbf{Y}_r), the minimal number of independent recordings J should be

not less than the number of sensors ($M = 24$), thus for each range gate $J > M$, (i.e. $J > 24$) independent realizations are to be recorded for each range gate $r = 1, \dots, R_r = 165$. In the case of $J < 24$, the data covariance matrix is rank-deficient; this means that if we apply the robust beamforming processing for sensor focusing, we inevitably will face the problem of huge artifacts (so called ghosts on the speckle corrupted scene images). In the radar terminology [2], these artifacts (speckle and ghost targets) will inevitably increase the false alarm rate.

3. Low Resolution Stage

The general mathematical formalism of the problem at hand and the DEDR framework that we employ in this section is similar in notations and structure to those in the previous studies [4], [5], [6], and some crucial elements are repeated for convenience of the reader.

3.1 Problem Formalism

The mathematical model of the power spectrum distribution (the so-called spatial spectrum pattern, SSP) restoration problem is stated as follows. Consider the unknown continuous spatial distribution of the extended radiating source within the given spatial domain (interval of analysis) $\Theta \ni \theta$ defined by the instantaneous complex amplitudes $e(t, \theta)$ of the source. In a convenient discrete-form representation, we consider the discretized interval of spatial analysis Θ with a set of K prescribed spatial directions $\{\theta_k; k = 1, 2, \dots, K\} \in \Theta$. The vector

$$\mathbf{e}(t) = \text{vec}\{e_k(t) = e(t, \theta_k); k = 1, \dots, K\} \quad (1)$$

composed with the complex amplitudes of the source signals from all K spatial directions is referred to as the vector of

random unknown instantaneous source complex amplitudes. For the Y- configured MIR at hand, the phase at the m th antenna element as a result of the k th source is $\omega_k y_m$ where $\omega_k = 2\pi \sin(\theta_k)$ and y_m is the location of the elemental phase center with respect to the midpoint of the array in wavelengths (λ). We assume the statistically uncertain scenario where t th time sampled signal at the m th element of the array for a fixed range gate r is

$$u_m(t) = \sum_{k=1}^K e_k(t) g_m(\theta_k) \exp(i\omega_k y_m) + n_m(t) \quad (2)$$

where $g_m(\theta_k)$ is the pattern response of the m th array element in the direction θ_k and $n_m(t)$ is the t th sample of the adopted Gaussian noise from the m th array element. This noise component is modeled as a random variable independent of both time index t and the element index m . The previous equation of observation (2) can be put in the following vectorial form:

$$\mathbf{u}(t) = \mathbf{S}\mathbf{e}(t) + \mathbf{n}(t); \quad t = 1, \dots, T \quad (3)$$

where $\mathbf{n}(t)$ represents the observation noise and \mathbf{S} is the signal formation matrix (SFO) defined as

$$\mathbf{S} = \text{matrix}\{S_{mk} = g_m(\theta_k) \exp(i\omega_k y_m)\} \quad (4)$$

where $m = 1, \dots, M$; $k = 1, \dots, K$. In (3), $\mathbf{u}(t)$, $\mathbf{e}(t)$ y $\mathbf{n}(t)$ represent zero-mean complex vectors composed of the sample coefficients $\{e_k(t), n_m(t), u_m(t); k = 1, \dots, K; m = 1, \dots, M\}$. These vectors are characterized by the correlation matrices [4]

$$\mathbf{R}_u = \langle \mathbf{S}\mathbf{R}_e\mathbf{S}^+ \rangle + \mathbf{R}_n \quad (5)$$

where

$$\mathbf{R}_e = \mathbf{D}(\mathbf{b}) \quad (6)$$

is the diagonal matrix with the vector-form SSP \mathbf{b} at its principal diagonal $\{b_k = b(\theta_k) = \langle e_k(t)e_k(t)^* \rangle = \langle |e_k(t)|^2 \rangle; k = 1, \dots, K\}$ and $\mathbf{R}_n = N_0\mathbf{I}$, respectively, where \mathbf{I} defines the identity matrix, N_0 is the observation white noise power, superscript $+$ stands for Hermitian conjugate and $\langle \cdot \rangle$ defines averaging.

3.2 DEDR-MSF Method

The low resolution RS imaging problem is stated generally as follows: to form the image of the tag $\hat{b}(\theta_k)$ as a function of the spatial scene coordinates applying the MSF method [4], i.e.

$$\hat{b}(\theta_k) = \mathbf{s}(\theta_k)^+ \mathbf{Y}_r \mathbf{s}(\theta_k); \quad k = 1, \dots, K \quad (7)$$

in which the image is formed as an MSF estimate of the SSP distribution over the remotely sensed scene at a particular r th range gate. In the pursued here non-parametric problem treatment, the resolution quality is assessed by the shape of the resulting system PSF associated with the MSF image (7)

of a single point-type target located at the scene origin at the corresponding range gate $r \in R$. In particular, the desired system PSF is associated with the shape that provides the lowest possible side lobes (and grating lobes) level balanced over the minimum achievable effective width of the PSF main beam [2].

In (7), $\mathbf{s}(\theta_k)$ is the k th array steering vector composed of the corresponding k th row ($k = 1, \dots, K$) of the regular SFO matrix \mathbf{S} and the estimate \mathbf{Y}_r (M -by- M) of the array spatial correlation matrix is computed via performing averaging over the J snapshots as:

$$\mathbf{Y}_r = \{\hat{\mathbf{R}}_u\} = \frac{1}{J} \sum_{j=1}^J \mathbf{u}_j \mathbf{u}_j^+ \quad (8)$$

Based on (7), let us next analyse the PSF of the MIR imaging system attainable with the employment of the conventional GeoSTAR-configured Y-shaped array. In Fig. 2, we present the PSF related to the MSF-based single target (TAG) imaging procedure (7) employing the GeoSTAR-configured Y-shaped sensor array radar. The PSF cross-section in the x-y imaging scene provide explicit information on the spatial resolution cells achievable with such configured imaging sensor array that employ the conventional 2-D MSF method (7) for RS image formation. The PSF in Fig. 2 is presented with an inter-element spacing $d_A = 2\lambda_0$, i.e., equal to the double of the carrier wavelength. Note that the most important characteristics of this PSF is the width of the main beam and the maximum level of the secondary lobes (including the suppressed grating lobes).

The next feature enhanced RS imaging problem at hand is to develop the framework (in this study, the DEDR-RAB method) and the related technique(s) for high-resolution estimation (feature-enhanced reconstruction) of the SSP \mathbf{b} , we tackle this situation in the next section.

4. DEDR-RAB Technique

The classical robust adaptive MVDR method [10] adapted for the high-resolution spatial spectrum pattern estimation as a solution to the problem $\hat{\mathbf{b}} = \text{est}_{\text{MVDR}}\{\mathbf{b}|\mathbf{u}\}$ results in the non-linear solution dependent strategy [10]:

$$\hat{b}(\theta_k) = \frac{1}{\mathbf{s}^+(\theta_k) \mathbf{R}_u^{-1} \mathbf{s}(\theta_k)}; \quad k = 1, \dots, K \quad (9)$$

optimal (in the MVDR sense) for the theoretical model-dependent covariance matrix inverse \mathbf{R}_u^{-1} . In the practical RS target imaging scenarios, the unknown exact (model) covariance matrix is substituted by its J -sample maximum likelihood estimate (8) that results in the corresponding MVDR estimation algorithm [1], [10]

$$\hat{b}(\theta_k) = \frac{1}{\mathbf{s}^+(\theta_k) \mathbf{Y}_r^{-1} \mathbf{s}(\theta_k)}; \quad k = 1, \dots, K \quad (10)$$

feasible for the full rank \mathbf{Y}_r only. From Bayes Minimum Risk Estimation Strategy [5], [6] and from simple algebra,

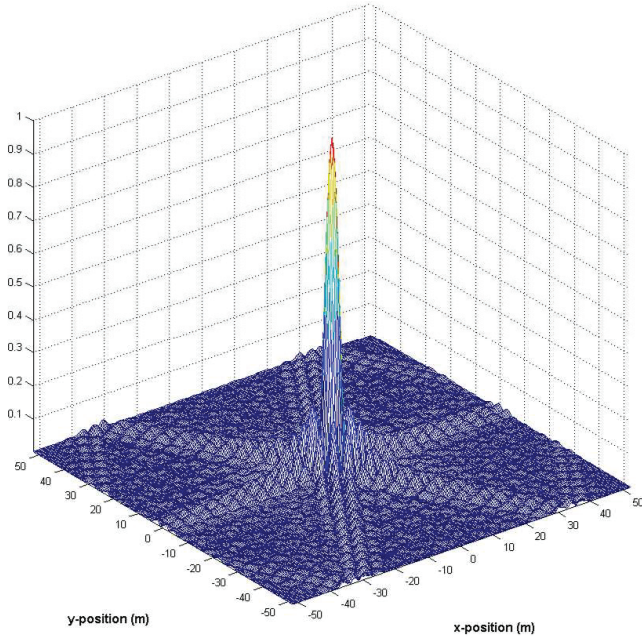


Fig. 2: Point Spread function (PSF) for 24 element Y-shaped configured multisensor imaging radar with $d_A = 2\lambda_o$ inter-element spacing for 30m range gate.

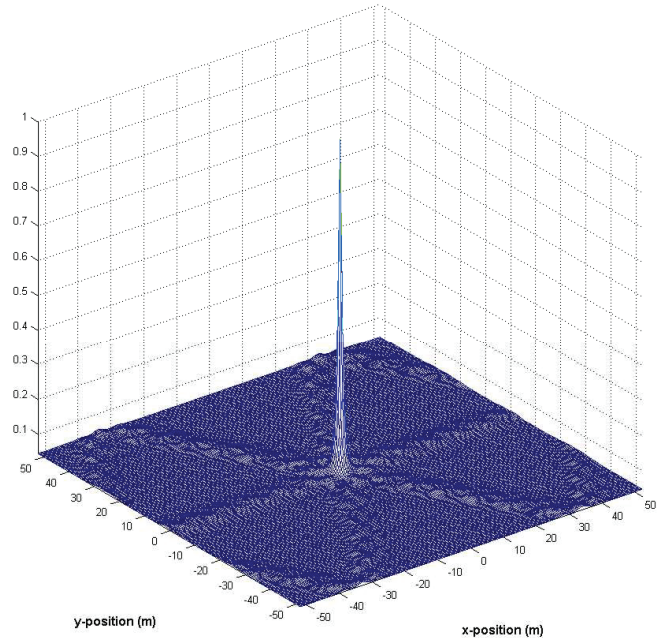


Fig. 3: Point Spread function (PSF) for 24 element Y-shaped configured multisensor imaging radar with $d_A = 2\lambda_o$ inter-element spacing for 30m range gate.

it is easy to corroborate that the theoretical model-based strategy (10) is algorithmically equivalent to the solution with respect to \mathbf{b} of the non-linear equation

$$\mathbf{D}(\hat{\mathbf{b}}) = \{\mathbf{W}(\hat{\mathbf{b}}) \mathbf{Y}_r \mathbf{W}^+(\hat{\mathbf{b}})\}_{\text{diag}} \quad (11)$$

with the solution operator (SO)

$$\mathbf{W}(\hat{\mathbf{b}}) = \mathbf{K}(\hat{\mathbf{b}}) \mathbf{S}^+ \mathbf{R}_n^{-1}. \quad (12)$$

Since $\mathbf{R}_n = N_0 \mathbf{I}$, the last SO becomes:

$$\mathbf{W}(\hat{\mathbf{b}}) = \mathbf{K}(\hat{\mathbf{b}}) \mathbf{S}^+ \quad (13)$$

where $\mathbf{K}(\hat{\mathbf{b}}) = (\Psi + N_0 \mathbf{R}_e^{-1})^{-1}$ and $\Psi = \mathbf{S}^+ \mathbf{S}$ represents the matrix-form PSF of the MSF low-resolution image formation system [2].

The DEDR framework [5], [6] suggests the worst case statistical performances optimization approach to the problem of $\hat{\mathbf{b}} = \text{est}_{\text{MVDR}}\{\mathbf{b}|\mathbf{u}\}$ with the model uncertainties regarding the statistics of the SFO perturbations that yields the robust (13). The RAB modification of the DEDR (DEDR-RAB) is constructed by replacing in (13) N_0 by the composite (loaded) $N_\Sigma = N_0 + \beta$. The latter is the observation noise power N_0 augmented by factor $\beta \geq 0$ adjusted to the regular SFO Loewner ordering factor and the statistical uncertainty bound for the for the SFO perturbation (see [5] for details). Finally, we can define the new DEDR-RAB as:

$$\mathbf{D}(\hat{\mathbf{b}}) = \{(\mathbf{S}^+ \mathbf{S} + N_\Sigma \mathbf{R}_e^{-1})^{-1} \mathbf{S}^+ \mathbf{Y}_r \mathbf{S} (\mathbf{S}^+ \mathbf{S} + N_\Sigma \mathbf{R}_e^{-1})^{-1}\}_{\text{diag}} \quad (14)$$

In Fig. 3, we present the PSF related to the DEDR-RAB single target (TAG) imaging procedure (14) employing the GeoSTAR-configured Y-shaped sensor array radar. The PSF cross-section in the x-y imaging scene provide explicit information on the spatial resolution cells achievable with such configured imaging sensor array that employ (14) for RS image formation. As in Fig.2, the PSF in Fig. 3 is presented with an inter-element spacing $d_A = 2\lambda_o$. Note the difference between the two PSFs at hand, that is, the width of the main beam and the maximum level of the secondary lobes (including the suppressed grating lobes).

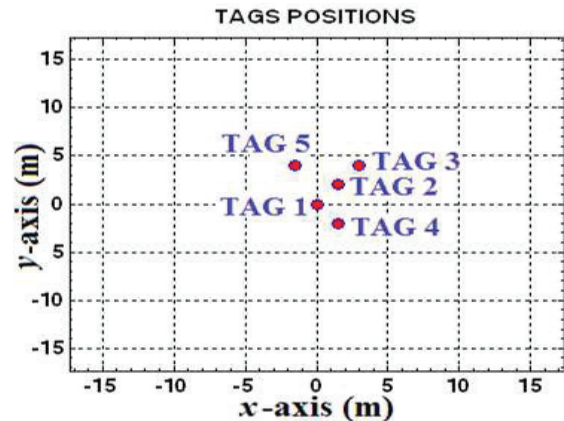


Fig. 4: Nominal multiple TAGs scene

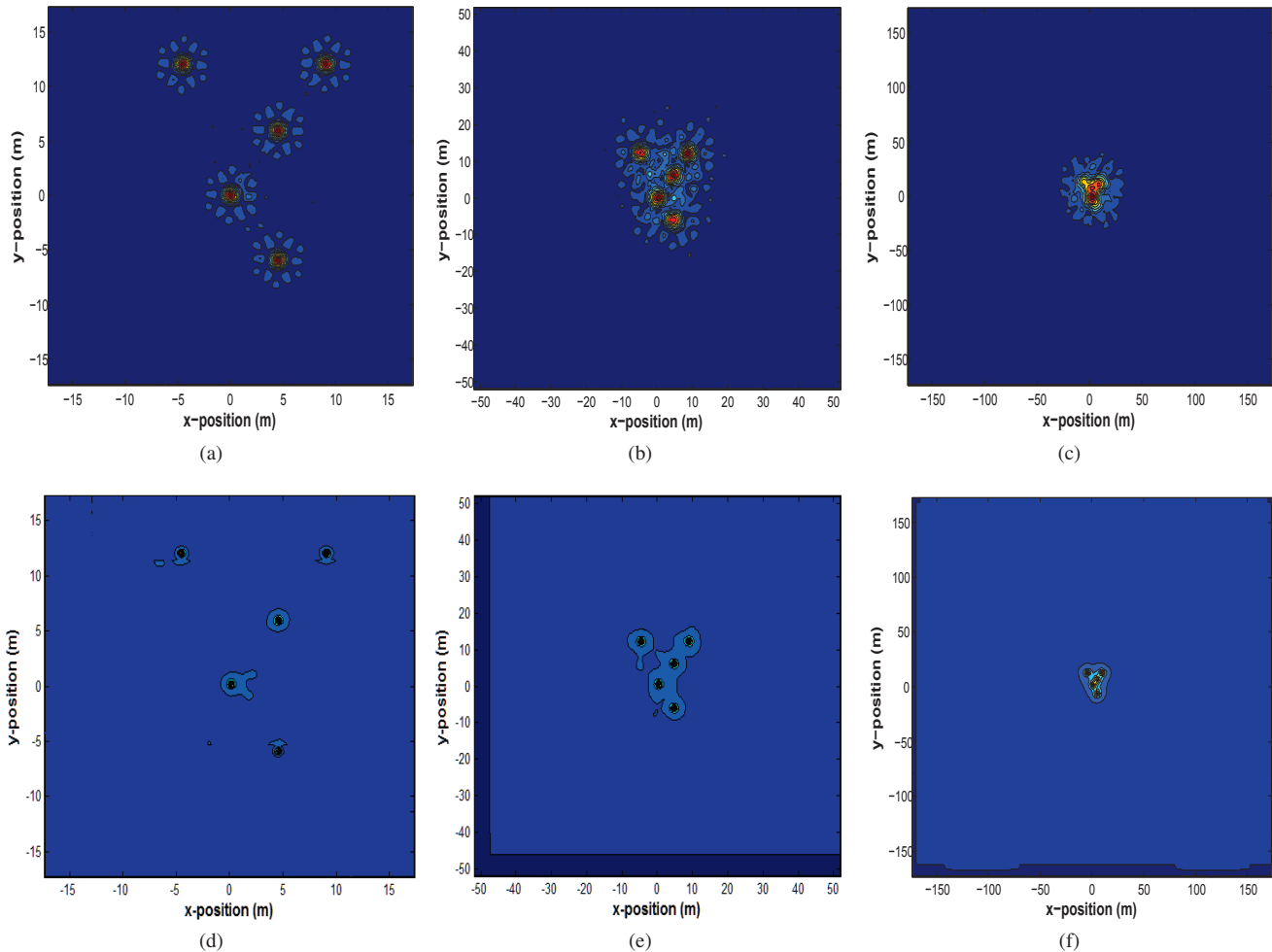


Fig. 5: Simulations protocols for the Y-configured MIR: (a)-(c) Low resolution scene image formed using the conventional DEDR-MSF technique (7), SNR = 10 dB, for a range gate of $r = 10\text{m}$, 30m and 100m respectively; (d)-(f) Feature enhanced image reconstructed employing the new aggregated DEDR-RAB method (14), SNR = 10 dB, for a range gate of $r = 10\text{m}$, 30m and 100m respectively.

5. Target Localization Protocols

We corroborated the effectiveness of the new DEDR-RAB technique (14) via simulation studies performed with the elaborated VRSL software. Three typical simulation protocols of radar imaging of a scene composed of five closely spaced targets (TAGs) in the range gates $r = 10\text{m}$, 30m and 100m (respectively, in columns) are presented in Fig. 5 for the Y-configured MIR. In Fig. 4 we present the nominal multiple TAGs scene. Figures 5a through 5c show the low resolution images of that scene formed using the conventional DEDR-MSF technique (7) for the 10 dB signal-to-noise ratio (SNR) typical for radar imaging scenarios [2] with a signal interference (INR) of 20 dB. Similarly, figures 5d through 5f present the feature enhanced (high-resolution) images of the same scene reconstructed with the proposed aggregated DEDR-RAB technique (14). The reported target

localization protocols are indicative of the drastically superior operational efficiency provided with method (14) that employs the DEDR-optimized GeoSTAR-configured array over other competing tested imaging array radar geometries [3].

To maintain consistency with the adaptive beamforming literature, we adopt the SINR as a measure of the effectiveness of our new aggregated DEDR-RAB method. The DEDR-MSF performs poorly against the new DEDR-RAB method as the latter presents an impressive average SINR of 30 dB compared to 12 dB for the DEDR-MSF when the signal of interest SNR=10 dB and exists a signal of interference of 20 dB. This is shown in Figures 6 and 7.

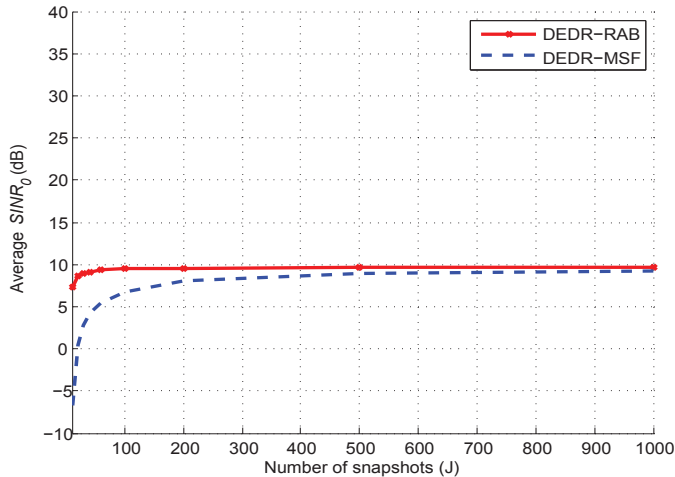


Fig. 6: Comparison of the DEDR-MSF and DEDR-RAB methods. Average SINR (dB) for a SNR=0 dB and INR=20 dB.

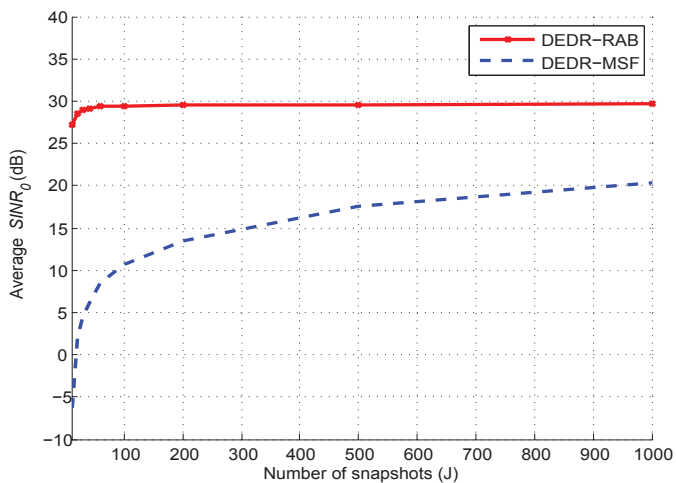


Fig. 7: Comparison of the DEDR-MSF and DEDR-RAB methods. Average SINR (dB) for a SNR=10 dB and INR=20 dB.

6. Conclusion

We have addressed the new robust DEDR-RAB approach for enhanced imaging of multiple target scenes in harsh operational environments directly adapted to MIR imaging systems, in this work particularly with Y-configured GeoSTAR MIR. The presented high-resolution target localization protocols are indicative of the superior operational efficiency of the Y-configured multimode imaging MIR system with the adopted GeoSTAR array geometry. The reported PSFs provide explicit information on the spatial resolution achievable with such MIR system that employs the proposed DEDR-RAB image formation technique. We demonstrated

via the analysis of behavior of SINR quality metric that method (14) yields the best imaging performances. In future studies, we intend to focus on the HW-SW co-design aimed at the resolution enhancement of the DEDR imagery and approaching the super-resolution imaging performances with MAR systems.

This will push forward our capabilities in the hardware-software codesign-based optimization of the RS and multi-sensor radar systems paving a way toward adaptive super-resolution sensing with the mm-waveband array radar systems.

References

- [1] J. Li and P. Stoica, *Robust Adaptive Beamforming*. John Wiley & Sons Ltd, 2006.
- [2] F. M. Henderson and A. J. Lewis, Eds., *Principles & Applications of Imaging Radar, Manual of Remote Sensing*, 3rd ed. John Wiley & Sons Ltd, 1998, vol. 2.
- [3] V. Espadas and Y. Shkvarko, "Descriptive experiment design framework for high resolution imaging with multimode array radar systems," *Applied Radio Electronics*, vol. 12, no. 1, pp. 157–165.
- [4] Y. Shkvarko, "From matched spatial filtering towards the fused statistical descriptive regularization method for enhanced radar imaging," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 16–16, Jan. 2006. [Online]. Available: <http://dx.doi.org/10.1155/ASP/2006/39657>
- [5] Y. V. Shkvarko, "Unifying experiment design and convex regularization techniques for enhanced imaging with uncertain remote sensing data, part i: Theory," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 1, pp. 82–95, Jan 2010.
- [6] —, "Unifying regularization and bayesian estimation methods for enhanced imaging with remotely sensed data-part ii: implementation and performance issues," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 5, pp. 932–940, May 2004.
- [7] A. Tanner, W. Wilson, B. Lambrigsten, S. Dinardo, S. Brown, P. Kangaslahti, T. Gaier, C. Ruf, S. Gross, B. Lim, S. Musko, S. Rogacki, and J. Piepmeier, "Initial results of the geostationary synthetic thinned array radiometer (geostar) demonstrator instrument," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 7, pp. 1947–1957, July 2007.
- [8] Y. Shkvarko and V. Espadas, "Experiment design framework for super-high resolution imaging with the geostar configured sensor array data," in *Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW), 2010 International Kharkov Symposium on*, June 2010, pp. 1–3.
- [9] Y. V. Shkvarko, "Estimation of wavefield power distribution in the remotely sensed environment: Bayesian maximum entropy approach," *Signal Processing, IEEE Transactions on*, vol. 50, no. 9, pp. 2333–2346, Sep 2002.
- [10] Y. Shkvarko, J. Tuxpan, and S. Santos, "Dynamic experiment design regularization approach to adaptive imaging with array radar/sar sensor systems," *Sensors*, vol. 11, no. 5, pp. 4483–4511, 2011. [Online]. Available: <http://www.mdpi.com/1424-8220/11/5/4483>

Archiving and visualization of the patient's anatomical model using B-spline curve and surface representation

A. Dominik Spinczyk

Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Silesia, Poland

Abstract—From the patient's perspective a minimally invasive surgery decreases trauma, the associated pain, and greatly reduces the post-operative recovery time, trauma and faster return to life. It has some disadvantages, like limited field of view and occluded anatomy, which could be overcome by registration of the preoperative patient's anatomical model to the patient's position in the operating room. The aim is archiving and presenting the patient's anatomical model in a manner which is easy to modification and integration into surgical workflow finally causes simplification in perception of operating field. B-spline curve and surface representation for anatomical models were selected. For creating the model from segmentation results, the global surface interpolation algorithm is applied. Scene graph oriented solution was proposed for 3D image data and anatomical model presentation. The whole scene graph is divided into independent branches, which accelerates rendering and simplifies data loading and building the scene graph process. Due to the lack of a "gold standard" for evaluation of the segmentation results obtained automatically, a manual way of correction is feasible. The presented approach was implemented in application for the anatomical liver model presentation using Open Inventor library.

Keywords: patient's anatomical model, operational field visualization, patient's anatomical model processing, surface visualization

1. Introduction

From the patient's perspective a minimally invasive surgery decreases trauma, the associated pain, and greatly reduces the post-operative recovery time, trauma and faster return to life [1]. In a minimally invasive surgery perception of operating field is one of the most important factors for successful intervention. Seitel et al. [2] proposed different methods of visualization of needle placement during ablation procedures indicating a significant influence on surgery accuracy. Bartoli et al. [3] indicate five disadvantages of laparoscopic approach: weak depth perception, constrained vantage point, limited field of view, weak haptics and occluded anatomy. There are a lot of efforts to overcome those limitations. Some of them could be overcome by using new hardware techniques as: 3D laparoscopic reconstruction, depth estimation or computer vision approach like Simultaneous Localization and Mapping (SLAM). It is very difficult to integrate a few hardware solutions together (eg stereo

camera and flexible head), so the combination of hardware and software solution is used. In the first step, the preoperative patient's anatomical model using 3D medical imaging like Computed Tomography (CT) or Magnetic Resonance (MR) is created. The 3D anatomical model is generated manually, semi-automatically or automatically. Due to the lack of a "gold standard" for evaluation of the segmentation results obtained automatically, a manual way of correction is feasible. After creating the model next important step is registration of the preoperative patient's anatomical model to the patient's position in the operating room [4]. Due to the absence of unambiguous mapping of coordinate systems during registration, this stage is often divided into two steps: first rigid registration and then deformable registration.

2. Materials and Methods

2.1 Model data and representation

Generally, the patient's anatomical model includes 3D medical images, usually in Digital Imaging and Communications in Medicine (DICOM) format and segmented anatomical structures. The model could be represented as point cloud, polygonal mesh, implicit surface equation or parametric equations [5]. The most common representation is point cloud (eg 3D Slicer [6] and Visualization toolkit [7]). In the presented approach, a parametric representation has been selected due to a few advantages [8]:

- parametric method on the plane can be easily extended to represent arbitrary curve in three-dimensional space,
- parametric curves feature a natural direction of traversal, where ordered sequences of points along a parametric curve can be generated and could be treated as the cross-section of the surface,
- parametric form is more natural for designing and representing shape because curve coefficients have a direct visual interpretation,
- parametric form can represent a surface by introduction of a second parameter.

The volume of stored data is smaller in a parametric representation than in point cloud case because only the net of control points and knots vector should be archived. Generally, B-spline surface $S(u, v)$ is a tensor product of B-spline curve, taking a bidirectional net of control points $P_{i,j}$, two vectors of knots u and v and the products of univariate

B-spline functions [8]:

$$S(u, v) = \sum_{i=0}^m \sum_{j=0}^n N_{i,p}(u)N_{j,q}(v)P_{i,j}, \quad (1)$$

where $N_{i,p}(u), N_{j,q}(v)$ denote B-spline basic functions. In the presented approach cubic B-spline functions are used ($N_{i,p}(u) = N_{i,4}, N_{j,q}(v) = N_{j,4}$).

2.2 Model creation

Anatomical model usually represents the surface of segmented structures. As regards surface fitting algorithm for labelled segmentation volume there are two approaches: approximation and interpolation [9]. In the interpolation case, which is used in the presented approach, a surface which passes directly through a given net of data points $D_{c,d}$ is constructed. All data points, which are necessary input data for interpolation algorithm, are found at the boundaries of a segmented region. Boundaries of a segmented region which is represented as a volumetric binary mask are found using Moore-Neighbor Tracing algorithm modified by Jacob's stopping criteria [10]. After finding, boundary points are treated as input data. The number of selected points depends on interpolated shape (eg for liver surface forty points for DICOM slice are used) and this is a compromise between model size and its accuracy. Vectors of knots are found assuming a uniform distribution of knots. To find unknown spline curve coefficients, surface global interpolation algorithm is calculated [9]. Surface global interpolation algorithm can be regarded as a generalization of curve fitting interpolation algorithm, which allows to find the curve of degree d passing through a set of points D_c in the order given. For the given net of data points $D_{c,d}$ and values of parameters $u = s_c$ and $v = t_d$ the corresponding data point satisfies the equation of the surface:

$$D_{c,d} = S(s_c, t_d) = \sum_{i=0}^m \sum_{j=0}^n N_{i,p}(s_c)N_{j,q}(t_d)P_{i,j}. \quad (2)$$

The input of the algorithm are net of data points, two knots vector and degree of B-spline curves. For fitting surface given a grid of $(m+1) \times (n+1)$ data points D_{ij} ($0 \leq i \leq m$ and $0 \leq j \leq n$) and a degree (p, q) , allows to find a B-spline surface of degree (p, q) defined by $(m+1) \times (n+1)$ control points that passes all data points in the given order. Surface interpolation algorithm has been used to find liver surface and curve interpolation algorithm for central lines of segmented vessels respectively.

2.3 Visualization application architecture

The main challenge for the imaging navigation system is presentation of a real-time model of the operating field, which comprises: a personalized model of the patient's anatomy, models of operational tools, additional components to facilitate an orientation in the space (coordinate system

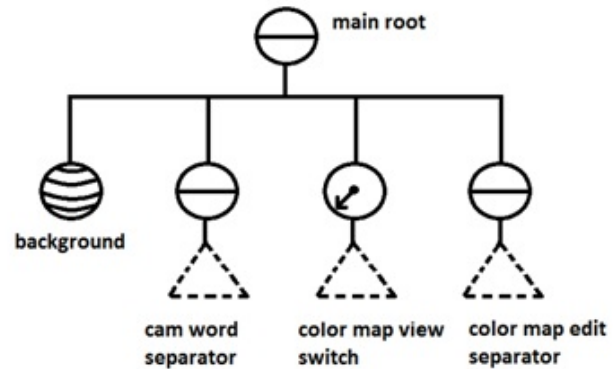


Fig. 1: Division of main scene graph in independent sub-branches

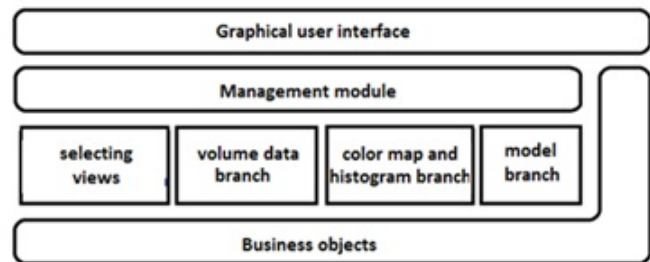


Fig. 2: Application architecture modules

orientation, volume data histogram, bounding boxes of the anatomical models, etc). Another important problem is how to store the data in a video card pipeline, mainly because of the size of personalized model. The most commonly used solution is scene graph [11], where the data objects in the rendering pipeline are represented only once in the graph and connected with references. The graphics engine renders the scene using recursive references while bypassing graph objects. Because of the need for real-time rendering, scene graph division is proposed to store individual components as independent scene graph sub-branches (cf. Fig. 1). Scene graph is presented in the figures in the form of the nodes with function description.

Application architecture (cf. Fig. 2) is based on a component approach [12], where graphical user interface includes scene graph viewer, management module encapsulates root of scene graph and represents interface of scene graph to the other application components. Every sub-branch of scene graph is implemented as a separate component.

The management module accepts commands from the user interface and triggers an appropriate events in the scene graph. It also takes care of creating the correct scene graph from loaded data. It includes classes responsible for the synchronization between the volumetric model (DICOM data), a model composed of the curves (boundaries of segmented

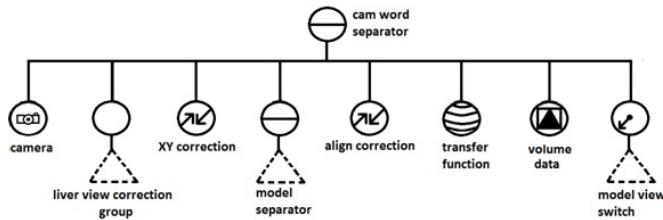


Fig. 3: Scene graph volume data branch

anatomical structures).

The colour map and the histogram components are responsible for enabling the user to change the colour to represent the voxels values that make up the volumetric image. Histogram presents a group of voxels of the same value in the entire volume data.

The selecting views component is responsible for providing functionality that allows the user to select a presentation manner, such as clipping volumetric model display, presenting it as a cross-section, or presenting the whole scene graph. Nodes in this module which are related to the editing mode, will be described in the next paragraphs.

In order to avoid unnecessary complexity in building a scene graph, the responsibility for its construction was divided into individual modules. This will allow the gradual creation, depending on the data being read with your selections, and the integration will take place in the layer of management.

2.3.1 Visualization Volume data

The sub-branch responsible for displaying volumetric data consists mainly of volume data node (cf. Fig. 3) which represents DICOM image data in the form of three-dimensional texture. Presenting manner of volume data depends on the selected visualization mode (cf. Fig. 4).

2.3.2 Visualization anatomical models

The model is presented in the form of spline curves. The main separator contains as many child nodes as many curves exist in the loaded model (cf. Fig. 5). Each curve defines the colour by which it is represented ("material" node), then a special node that stores the control points and knots appropriate to display curves ("coordinates" node).

For each curve a separator node is defined that holds the points that appear on the screen, which allows a user to edit interactively the coordinates of the selected control point, by selecting and moving it (cf. Fig. 6). The single control point is represented as a sphere node, manipulator and sensor. The manipulator is presented as a cube. The manipulator movement is detected by the associated sensor and it updates the previous position ("delta translation" node). The translation shall be placed before the ball, causing it to move in sync with the manipulator.

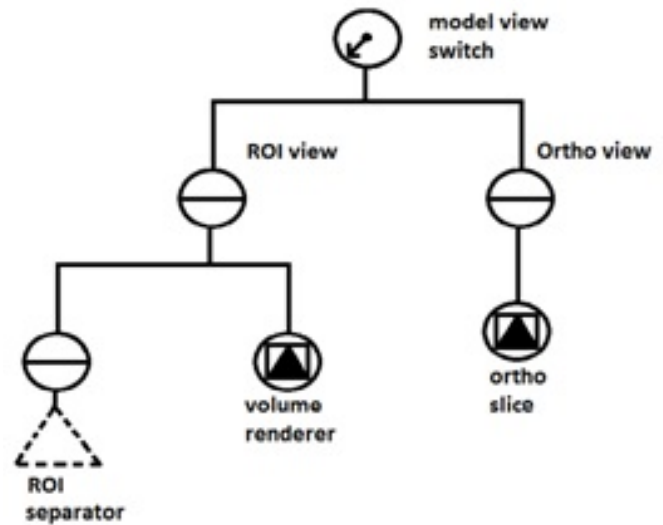


Fig. 4: Switching visualization mode of volume data

3. Experiments and evaluation

The presented approach was implemented in the application for anatomical liver model presentation using Open Inventor library [11] and have been evaluated using the data set consisted of twenty abdominal CT obtained from the Department of Clinical Radiology of Silesian Medical University. The graphical user interface is presented in the next figures (cf. Fig. 7, 8), where selected structures: liver surface and vessels, are presented in green colour against DICOM data presented in a grey-scale map. The whole model of the liver and tumours or selected cross-section could be presented against DICOM data and a manual correction is possible.

In general, the model verification process is a complex issue. All selected and segmented structures should be validated separately. To simplify this process, the advantages of the parametric model representation is applied. As mentioned earlier, cubic spline curves are used which are characterized by the local control property. Moving the control point corresponding to the B-spline curve segment will only affect three local segments, which propagate shape changes to adjacent cross-sections (cf. Fig. 7). The individual segments of vascular structures are represented in the form of spline curves. The verification of results takes place in the volumetric visualization mode due to the variation in the direction of the vessels (cf. Fig. 8). It is also possible to insert additional control points into the curve if necessary. The verified anatomical model could be stored in data files.

4. Discussion and Conclusion

Representing an anatomical model in parametric curve and surface is useful. From archiving point of view it takes a

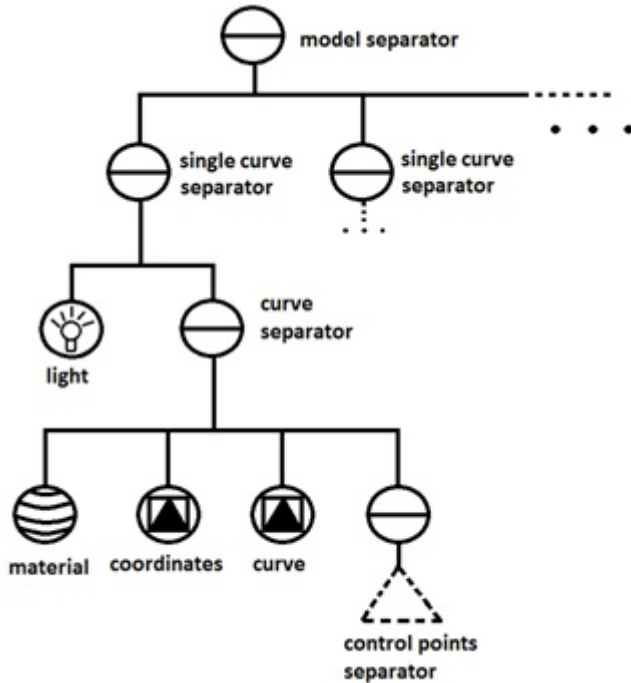


Fig. 5: Scene graph branch for model visualization

smaller place because it is not necessary to keep all surface point positions but only control points and knots vectors. From the presentation point of view, control points have important properties, so that moving them influence only the local segments. Smoothness of B-spline curves, depending on basic functions degree, could be applied to model anatomical shapes which are usually smooth. Additionally, the curve is included in a convex polygon stretched at control points. Visualizing control points and making them draggable make the whole curve draggable.

Parametric representation is more powerful for data traversal, finding cross-section, etc.. It is also possible to change internal representation of spline curve and surface from weighted form of B-spline basic function for pp-form [13], which is based on derivative matrix and is more effective for computation. Using the procedure for inserting a node parametric representation could be treated as a quasi multi-resolution model which could interpolate shape more precisely by inserting an additional segments. To fit more difficult shape it is also possible to change B-spline basic function degree.

The scene graph is useful technique for anatomical model presentation. Basically, it collect all rendering objects in an effective way, but using more mature implementation could be connected with collision detection module and notifies the collision occurrence of a surgical tool and important anatomical structures. The division of scene graph into independent branches (implementation of "divide and conquer" paradigm) and introduction of management module

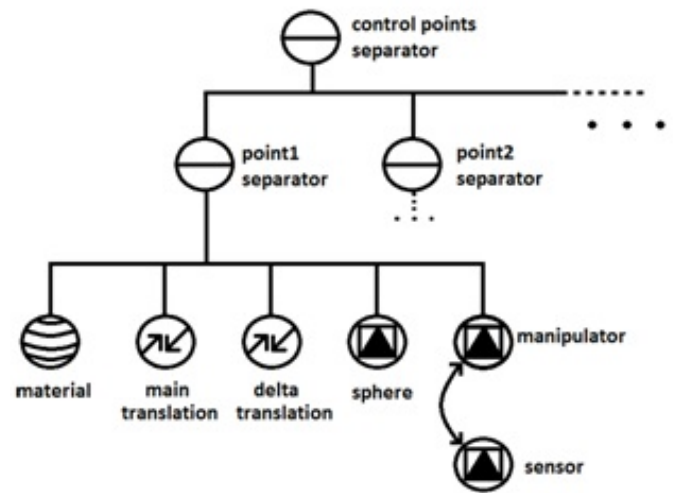


Fig. 6: Scene graph branch for manual curve correction

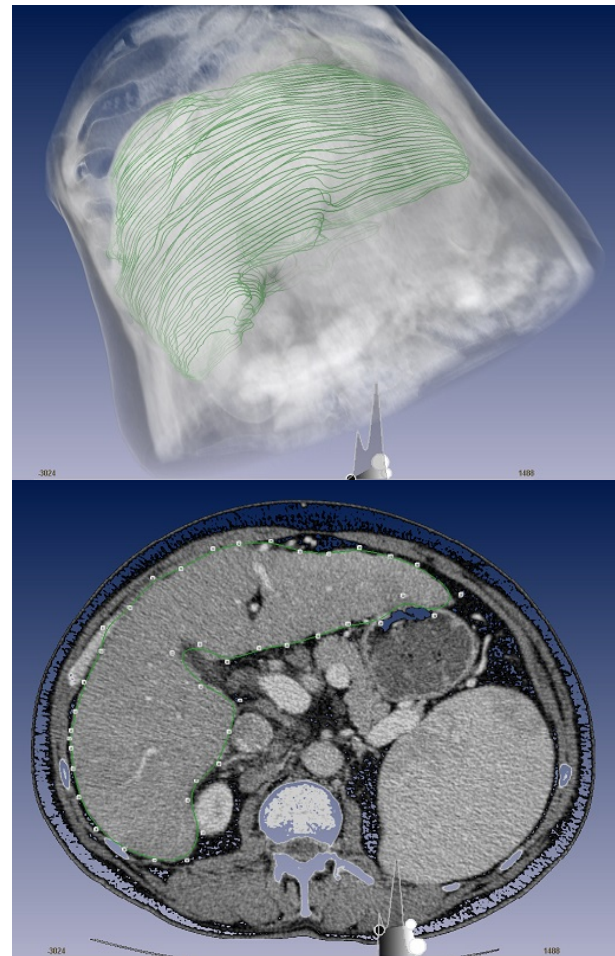


Fig. 7: Liver surface model representing as set of B-spline curves (up), and control points for selected curve for manual correction (down).

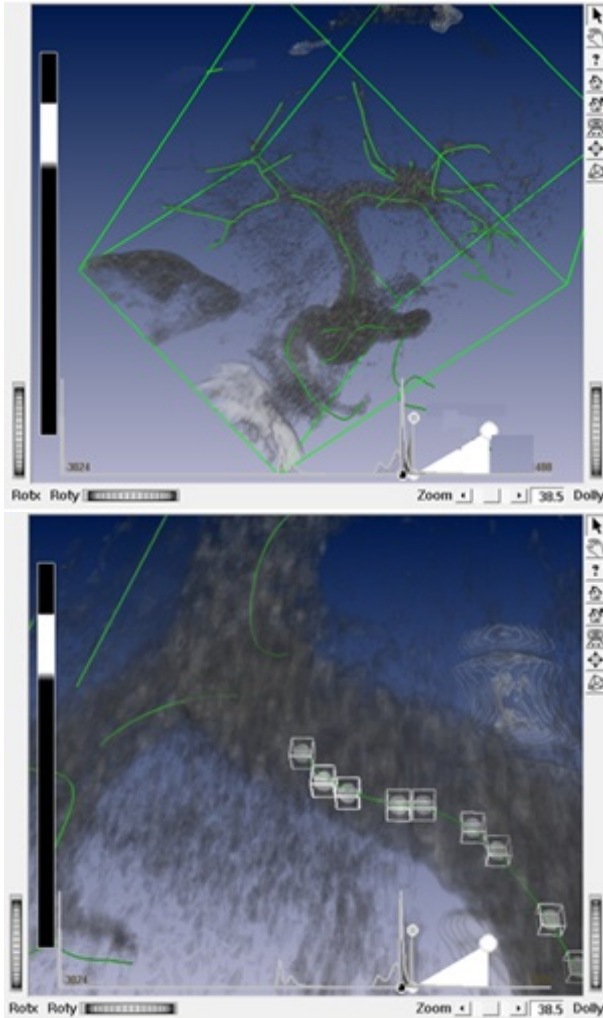


Fig. 8: Liver vessels model representing as set of B-spline curves (up), and control points for selected curve for manual correction (down).

simplify the integration of this technique for different clinical cases even for rendering sophisticated scenes.

From the global point of view usefulness of patient's anatomical model depends on the proper use in surgical workflow. To avoid mistakes, a close cooperation between the radiologist and surgeon is desirable. After manual correction of automated generated model in a planning phase, possible intervention techniques, are easier to compare (by selecting a target an an entry points and tracing the possible surgical tools trajectories). After an initial registration of the preoperative model and patient physical position, anatomical model could be useful during interventions. Parametric representation simplifies presentation only selected part of the model. In some planning intervention methodology key frame of DICOM slice is selected (eg frame corresponding to diaphragm position) to take into account deformation of anatomical structures [14] caused by breathing or emphasis

of laparoscopic instruments [15]. The values of B-spline model parameter for corresponding cross-section could be stored and the ability to visualize specific cross-section of the model is desirable.

5. Acknowledgment

The study was supported by National Science Center, Polad, Grant No UMO-2-12/05/B/ST7/02136.

References

- [1] M. Candiani, S. Izzo, A. Bulfoni, J. Riparini, S. Ronzoni, A. Marconi, Laparoscopic vs vaginal hysterectomy for benign pathology, *American Journal of Obstetrics and Gynecology* 200 (2009) 368.e1 - 368.e7.
- [2] A. Seitel, L. Meier-Hein, S. Schawo, B. Radele, S. Mueller, F. Pianka, B. Schmied, I. Wolf, H. Meinzer, In-vitro evaluation of different visualization approaches for computer assisted targeting in soft tissue, *Computer-Assisted Radiology and Surgery 2 Supplement 1* (2007) S188-S190.
- [3] A. Bartoli, T. Collins, N. Bourdel, M. Canis, Computer assisted Minimally Invasive Surgery: Is medical Computer Vision the answer to improving laparosurgery?, *Medical Hypotheses* 79 (2012) 858 - 863.
- [4] T. Peters, K. Cleary, *Image-guided interventions: technology and applications*, Springer 2008.
- [5] J. Foley, A. van Dam, S. Feiner, J. Hughes, R. Phillips, *Introduction to Computer Graphics*, Addison-Wesley 1994.
- [6] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, R. Kikinis, *3D Slicer as an image computing platform for the Quantitative Imaging Network*, *Magnetic Resonance Imaging* 30 (2012) 1323 - 1341.
- [7] W. Schroeder, K. Martin, B. Lorensen, *The Visualization Toolkit*, Pearson Education 2004.
- [8] L. Piegl, W. Tiller, *The NURBS Book*, Springer 1996.
- [9] P. Dierks, *Curve and Surface Fitting With Splines*, Clarendon Press 1993.
- [10] R. Gonzalez, R. Woods, S. Eddins, *Digital Image Processing Using MATLAB*, Pearson Prentice Hall 2009.
- [11] J. Wernecke, *The Inventor Mentor: Programming Object-Oriented 3D Graphics with Open Inventor*, Addison-Wesley 2002.
- [12] C. Szyperski, D. Gruntz, S. Murer, *Component Software - Beyond Object Oriented Programming*, Addison-Wesley 2002.
- [13] C. de Boor, *A practical Guide to Splines*, Springer 1978.
- [14] A. Hostettler, S. Nicolau, Y. Rmond, J. Marescaux, L. Soler, A real-time predictive simulation of abdominal viscera positions during quiet free breathing, *Progress in Biophysics and Molecular Biology* 103 (2010) 169 - 184.
- [15] D. Spinczyk, A. Karwan, J. Rudnicki, T. Wroblewski, Stereoscopic liver surface reconstruction, *Videosurgery and Other Miniinvasive Techniques* 7 (2012) 181 -187.

Mobile Videos Quality Measurements for Long Term Evolution (LTE) Network

Hamad Almohamedh, Fahad Al Qurashi, Ivica Kostanic

Department of Electrical and Computer Engineering

Florida Institute of Technology

Melbourne, Florida, USA

halmoham@my.fit.edu, falqurashi2008@my.fit.edu, kostanic@fit.edu

Abstract - This paper presents an evaluation of live mobile video streaming measurements over a Long-Term Evolution (LTE) network using User Datagram Protocol (UDP). Also, it describes a high quality videos database that was created as a part of the evaluation. The objectives are to quantify the impact of Receive Signal Strength Indicator (RSSI), Reference Signal Received Power (RSRP) and Reference Signal Received Quality (RSRQ) measurement on mobile video streaming over an LTE network and to provide 4k resolution raw videos for video quality assessment. A testing environment is created with a client server method to record parameters. For testing of video streaming, a real AT&T network is used at multiple locations for the same video to get different readings. Results of the live video streaming and live measurement show high consistency and correlation between the RSSI, RSRP, RSRQ and packet loss and provide the observation that these parameters vary at different locations within the cellular network

Keywords: LTE; Mobile Video Streaming Quality; RSRP; RSRQ; RSSI

1 Introduction

Broadband cellular access technologies and smart phones have brought about a major increase in the traffic over cellular networks. Data centric applications have been gaining importance. Among them the video streaming is one of the most resource consuming one. This has played a major role in tremendous increase in the traffic. Based on Cisco's predictions, cellular video streaming traffic in the year 2014 will be double that of the previous year 2013. See figure 1 [1].

AT&T projects growth of mobile video traffic in the range of 8000% over the 4-year period between 2011 and 2014. Cellular device manufacturing companies predict that this growth will continue to increase at an average rate of 92% per year for the next 5 years [2]. Different studies of mobile video usage show that with the increase of fast and handy devices, such usage continues to increase to even greater volumes. Different studies on the mobile video usage have shown that with the increase of fast and handy devices, the mobile video streaming has increased to a great volume. Easily adaptable and faster cellular networks have

contributed towards mobile video steaming explosion. People are getting addicted to it as they can access videos anytime and anywhere from their mobile phones. Most studies have focused on the need for improvement of quality of the videos and design of interface for the video browsing [3].

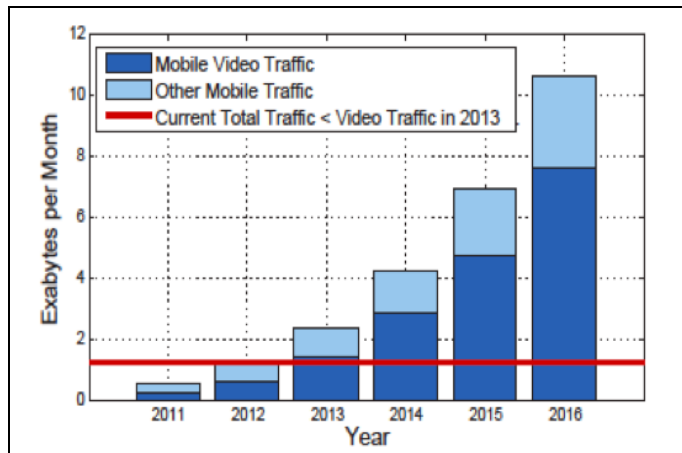


Figure 1. Prediction of mobile video traffic taken from [1].

2 VIDEO QUALITY AFFECTING FACTORS OVER LTE NETWORKS

Video quality over the LTE network is dependent upon several factors that contribute to overall quality. It is important to note some of these key factors that affect streaming video quality over LTE cellular network.

2.1 Reference Signal Strength Indicator (RSSI)

RSSI is the total power received by the resource element in dBm. The resource element is the smallest unit which consist of one subcarrier for duration of one orthogonal frequency division multiplex (OFDM) symbol [4]. RSSI is a combination of the signals received from all sources including the power from serving cell, non-serving cell, co-channel, and adjacent channel interference [5].

2.2 Reference Signal Received Power (RSRP)

RSRP is the reference signal received power that measures the power in a single resource element. RSRP main purpose is to help determining the serving cell for initial random access or LTE handover. RSRP value rang from -140 to -44 dBm [6].

2.3 Reference Signal Receive Quality (RSRQ)

RSRQ is the quality of the power received in the resource element. RSRQ is a measure of the quality of the signal rather than the quantity of received signal strength. RSRQ is dependent on both RSSI and RSRP and can be calculated by using the following equation (1) where N is the number of resource blocks (RBs) used for RSSI.

$$RSRQ = N \frac{RSRP}{RSSI}. \quad (1)$$

2.4 Packet Loss

Data communication takes place in the form of packets. Packet loss results when packets fail to reach the desired destination [7]. As such, the packet loss represents a fundamental measure of the quality for a data communication link. In LTE networks, the packet loss is related to factors such as RSSI, RSRP and RSRQ.

3 EXPERIMENT METHODOLOGY

Setting up the experiment in a simple and professional manner is important for the testing methodology. The main aim is to test video quality over the LTE network using simple devices that can manage the task. User equipment is registered with the LTE base station. This base station is further connected to the server through the Internet as shown in figure 2.

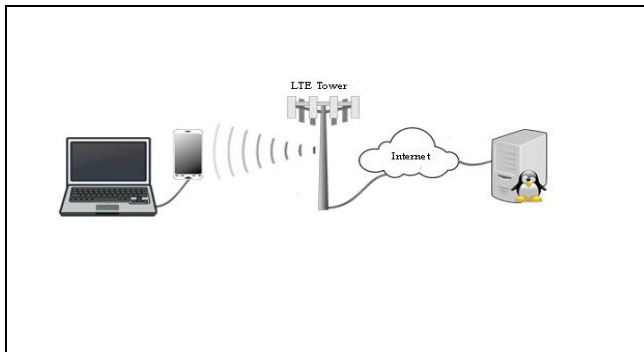


Figure 2. Experiment Process Illustration

In all test cases, Samsung Galaxy S3 is utilized as user equipment (UE) in a real AT&T LTE network. All

measurements were performed in 10MHz LTE network in Melbourne, Florida, USA. Video traffic was streamed over UDP protocol using a mobile video quality prediction (MVQP) streaming server. A Samsung Galaxy S3 is connected to a laptop and signal measurements are stored in both the laptop and in the MVQP server. The streaming video process starts with shooting the raw videos which are then converted to MP4 and down sampled to flow in the LTE network.

3.1 SOURCE VIDEOS

The videos used in this study were recorded with a Sony PMW-F5 camera. A 16-bit RAW data were captured at a resolution of 4K. This resolution preserves more tonal values than what may be differentiated by human eye. All videos recordings were captured at 29.97 frames per second in various selected locations in city of Melbourne, Florida, USA. A total of 10 videos were used in this study from a collection of 40 videos that are available in the MVQP database [8]. Figure 2 provides a brief description for each of the ten videos [9].

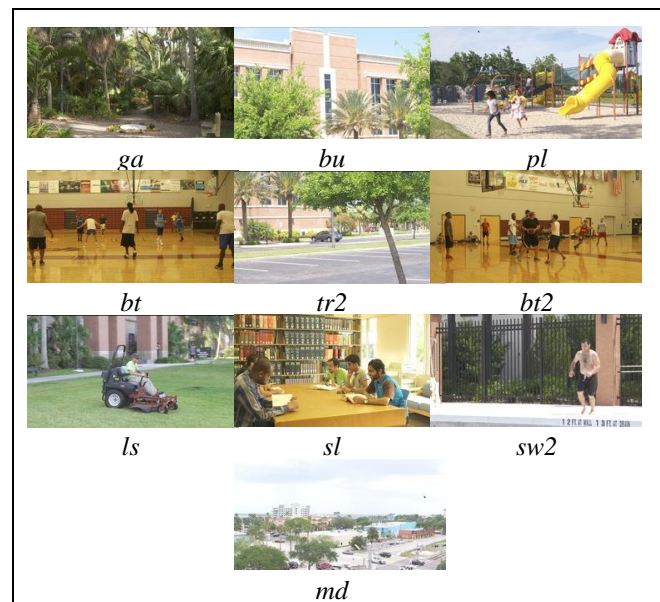


Figure 3. Thumbnail for each video frame.

- 3.1.1 Garden (ga).** Shot at Florida Institute of Technology (FIT) campus garden on a sunny afternoon. There are light contrasts. The camera tilts the trees from bottom to top.
- 3.1.2 Building (bu).** Shot at FIT's campus on a sunny afternoon. The building is surrounded with number of trees. The camera pans from left to right.
- 3.1.3 Playground (pl).** Shot in a park on a sunny afternoon. Children are playing on slides. Filled with bright and fascinating colors.

- 3.1.4 Basketball Training (bt).** Shot inside Clemente Center at FIT. Many players show diverse contrasts and complex motions. The camera is stationary.
- 3.1.5 Tree (tr2).** Shot near a road side on a sunny afternoon. The camera pans across the scene from up to down.
- 3.1.6 Basketball Training (bt2).** Shot inside Clemente Center at FIT. Different ratios of light are shown with the movement of players. The camera shows a steady movement.
- 3.1.7 Lawn Service (ls).** Shot at FIT campus. A man is mowing a lawn. The camera tracks him from left to right.
- 3.1.8 Students at Library (sl).** Shot in main library at FIT on early morning. The stationary camera zooms out.
- 3.1.9 Swimming Pool 2 (sw2).** Shot at FIT's swimming pool on a sunny afternoon. A man jumps into the swimming and bright twinkle of waves are clearly visible in water. The camera pans from right to left.
- 3.1.10 Melbourne Downtown (md).** Shot from the top of the roof on a cloudy afternoon. The entire area is comprised of tall buildings and trees and various cars are moving on the road. The camera pans from right to left.

3.2 Server/Client Setup

The server/client were set up for live video streaming. The server used in this study is based on central processing unit of Intel, Core i3, 2100 with memory of 8 GB. Graphics card is Gallium, 0.4 on AMD CEDAR. The operating system is Linux based Ubuntu 13.10 working on 64-bit version. The hard drive is 1 Terabyte revolving at 7200 RPM. Gigabit Ethernet is used to incorporate speeds of 1 GB/s and Internet speed average is 256 Mb/s for the uplink and downlink. The client machine is a Lenovo ThinkPad L430 laptop working on a Core i3 processor of 2.4 GHz. The memory is same as the server, 8 GB. Hard drive is 240 GB solid-state drive. Operating system is Windows 7, 64 bit professional. It is set up to receive mobile video streaming from the server through Samsung Galaxy S3 connected with USB cable. FFMPEG tools [10] were installed in both the server and the client computer for streaming videos. The streamed videos were saved in the client computer to conduct the mean opinion score (MOS) in future work. FFMPEG is also used to measure streaming parameters such as packet loss, bit rate, and codec type which are then saved in file.

3.3 MVQP MEASUREMENT APPLICATION

MVQP measurement application is developed for the android platform compatible with the Samsung device that

works on the android operating system. MVQP application is used to measure and save radio frequency (RF) signals each second. This application saves the values of the parameters such as RSSI and RSRQ at each instant for later use and analysis.

3.4 LIVE MEASUREMENTS AND DISTORTED VIDEOS

Random locations were selected based on RSSI levels ranging from values of -87 dBm to -51 dBm. For each location, live streaming was done over the UDP protocol. The following steps were carried out in order to complete the study.

- 3.4.1** Streaming server was run by using FFserver for the selected videos.
- 3.4.2** FFMPEG was run in the client PC to receive and save the distorted video over the LTE network.
- 3.4.3** MVQP measurement application was started to save the Radio frequency signal measurement.

4 STUDY ANALYSIS AND RESULTS

Ten different locations were selected for this study. Figures 3-12 show the values of RSSI, RSRP, RSRQ, and the percentage of the total packet loss for each video in each location. The graphs illustrate that when RSSI, RSRP, and RSRQ decrease, the packet loss increases, affecting video quality based on the total percentage of lost packet. Chart legends show different colors for RSSI, RSRQ, RSRP and packet loss. The garden (ga) video's graph depicts a strong dip for RSSI RSRP and RSRQ values at locations 5 and 9 where there are peaks of packet loss (as shown in figure 3). RSSI values are less than -79 dBm, RSRP values less than -111 dBm, and RSRQ less than -12 dB at these locations where packet loss occurred. Remaining graphs indicate very similar in behavior, as shown below.

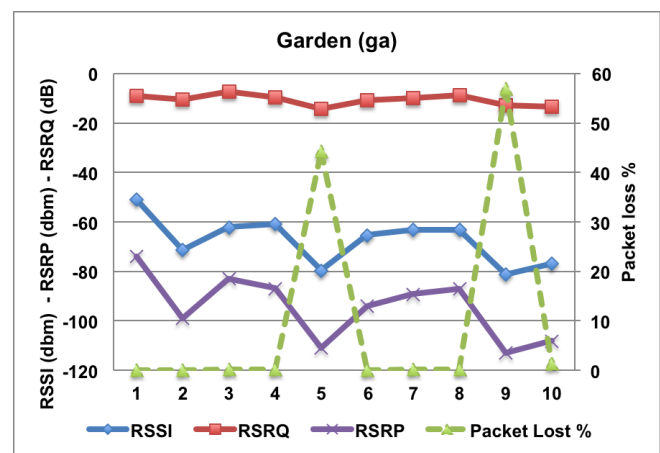


Figure 4. Garden (ga) video

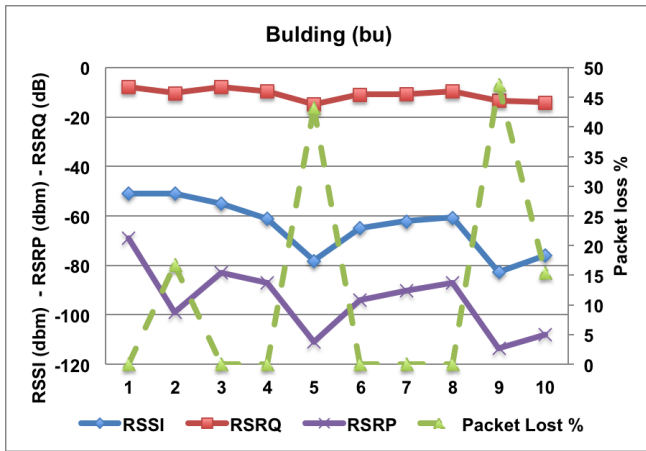


Figure 5. Building (bu) video



Figure 8. Tree 2 (tr2) video

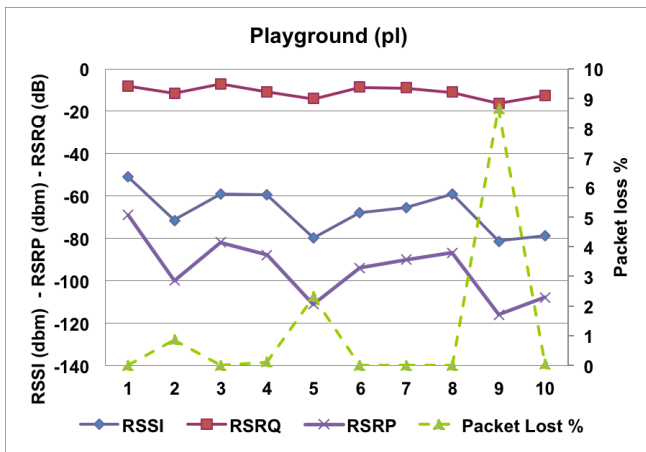


Figure 6. Playground (pl) video

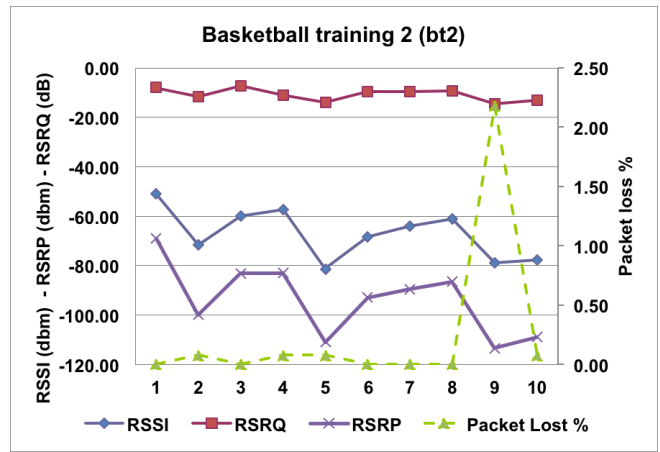


Figure 9. Basketball training 2 (bt2)

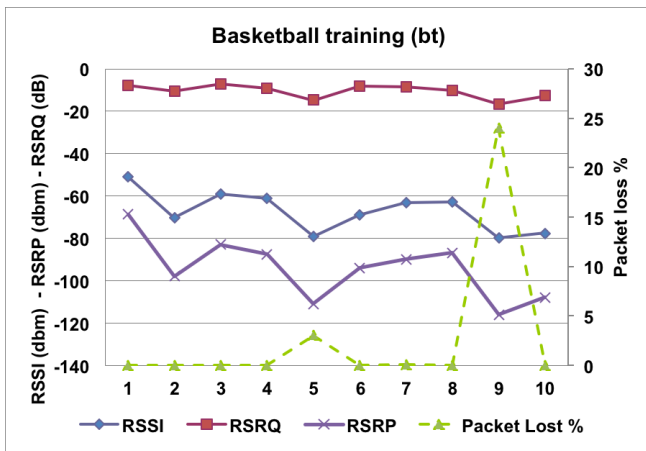


Figure 7. Basketball training (bt) video

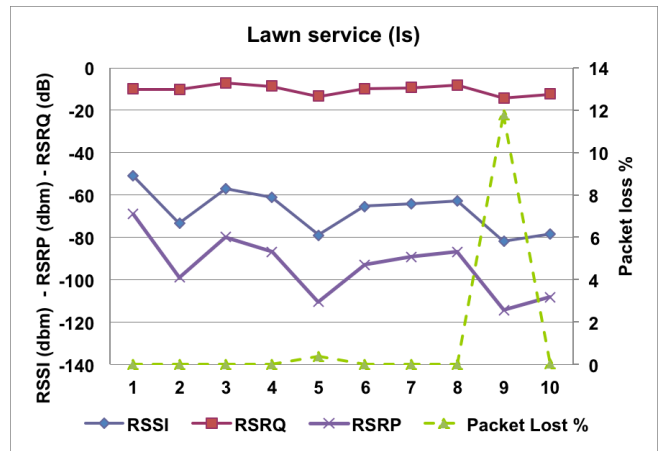


Figure 10. Lawn services (ls) video

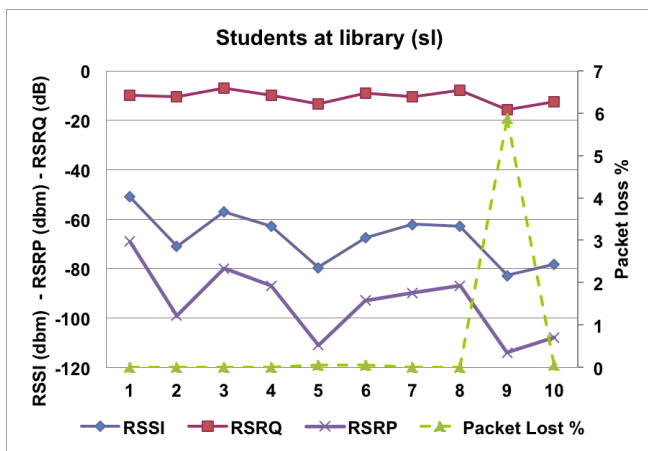


Figure 11. Students at library (sl)

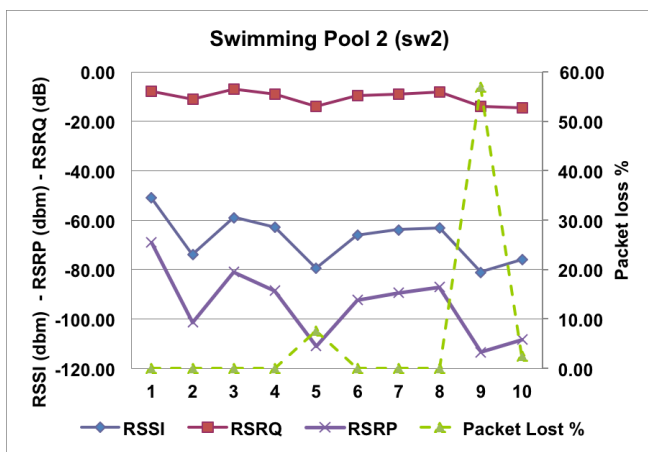


Figure 12. Swimming pool 2 (sw2) video

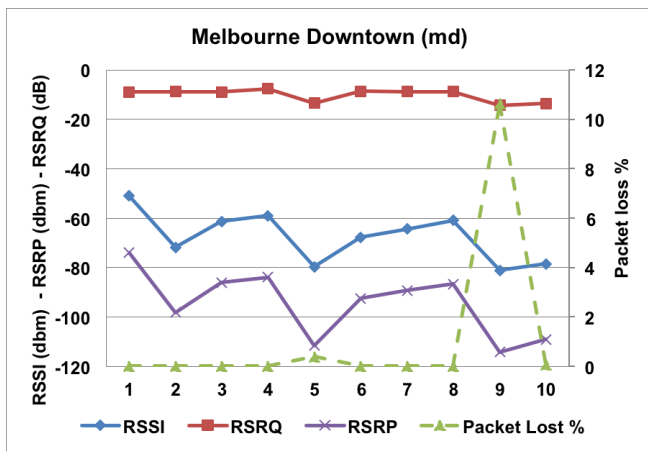


Figure 13. Melbourne downtown (md) video

5 CONCLUSION

In this experiment, three major factors are studied at 10 different locations for live streaming videos. Results of live measurement show high consistency and correlation between the signal strength, RSRQ, and packet loss. For RSRQ, lower values than -10 dB correspond to high packet losses in the video. The study also shows that locations 5 and 9 in particular are not suitable for video streaming, possibly due to congestion or higher distance from the serving cell. This implies that the quality of service varies at different locations with the cellular network. An extended study based on this paper was conducted. The study focused on compute the mean opinion score (MOS) and compared it with the packet loss [11].

6 References

- [1] Ayaskant Rath, Sanjay Goyal, and Shivendra Panwar, "Streamloading:Low Cost High Quality Video Streaming for Mobile Users", Polytechnic Institute of New York University, 2013
- [2] Jeffrey Erman, Alexandre Gerber, K.K. Ramakrishnan, Subhabrata Sen, and Oliver Spatscheck. "Over The Top Video: The Gorilla in Cellular Networks," AT&T Labs Research, New Jersey, USA.
- [3] Wei Song Dian, Tjondronegoro, and Michael Docherty, "Quality Delivery of Mobile Video: In-depth Understanding of User Requirements," Queensland University of Technology, Australia
- [4] Jack L. Burbank, Julia Andrusenko, Jared S. Everett, and William T. M. Kasch. 2013. Wireless Networking: Understanding Internetworking Challenges, First Edition, Wiley-IEEE Press, 2013.
- [5] Volkan Sevindik, Jiao Wang, Oguz Bayat, and Jay Weitzen, "Performance Evaluation of a Real Long Term Evolution (LTE) Network," in 8th IEEE International Workshop on Performance and Management of Wireless and Mobile Networks, Clearwater, Florida, 2012.
- [6] Ralf Kreher and Karsten Gaenger, LTE signaling, troubleshooting, and optimization, First Edition, Wiley, 2011.
- [7] Imed Bouazizi, "Estimation of packet loss effects on video quality," Control, Communications and Signal Processing, 2004. First International Symposium on , vol., no., pp.91,94, 2004
- [8] Fahad Al Qurashi, Hamad Almohamedh, and Ivica Kostanic, "MVQP Project, " Internet: <http://research.fit.edu/wice/mvqp.php>, Mar. 2014
- [9] Fahad Al Qurashi, Hamad Almohamedh, and Ivica Kostanic "RAW Video Database of Mobile Video Quality Prediction (MVQP)," (in-press), 2014
- [10] FFmpeg, Internet: <https://ffmpeg.org/index.html> .
- [11] Hamad Almohamedh, Fahad Al Qurashi, and Ivica Kostanic "Subjective Assessment of Mobile Videos Quality for Long Term Evolution (LTE) Cellular Networks," (in-press), 2014

Object Tracking Using SIFT and Kalman Filter

Seok-Wun Ha¹, Yong-Ho Moon²

^{1,2}Department of Informatics, ERI, Gyeongsang National University, Jinju, Rep. of Korea

Abstract - In computer vision and its related fields, an exact detection and an effective tracking of an object is very important. In the scale invariant feature matching between current and previous frame images by SIFT algorithm, a few of the matched keypoints are far away from each other in location, and therefore a location mismatching is caused and then a searching area is changed severely, and ultimately the tracking is defeated. By adding Kalman filtering to the feature matching, we resolved this defeating problem and got an effective object tracking performance. Experimental results show a more robust tracking ability in various and complicating cases.

Keywords: Object Tracking, SIFT, Kalman Filter

1 Introduction

In fields of computer vision and surveillance system, the exact and effective object tracking is necessary to guide a pathway or prevent a variety of criminals. For object tracking it needs a series of steps such as object detection, object area determination, invariant feature extraction, feature matching, and object tracking.

To detect and track a presenting object it needs to extract the invariant features and we used the Shift Invariant Feature Transform (SIFT) algorithm [1] proposed by D. G. Lowe. There are several algorithms that extract the invariant features from the specific target or image such as SIFT, GLOH [2], and SURF [3]. SIFT gives a robust characteristic for size, noise, brightness, and local distortion. GLOH shows the better performance for a structured image comparing with SIFT, but it has a disadvantage that the grid size for feature extraction extends to be doubled. SURF represents a rapid matching speed as compared with SIFT but SIFT is more useful than SURF in feature extraction of a small object.

In object tracking, because the absolute values of location and size for the comparable objects between two consecutive frames are altered, the size of the windows and their corresponding locations should be recalculated adaptively window [4]. But for the representatives among the matched keypoints their relative locations may be changed with large distance and, because of this, the object window is enlarged excessively and therefore the object tracking was be defeated ultimately. To solve this problem we utilized the Kalman Filter after detection of the representative keypoints [5]. Using several videos with a variety of movements of multiple objects the performance of the proposed system is experimented and

the tracking performance could be advanced effectively.

2 Object tracking system

The object tracking system is totally composed of four parts of obtaining the object-included window and the object area based on the multi-lateral histogram, extraction of the invariant features from the object-included window and the object area using SIFT algorithm, computing the matching rate and controlling the size of the object-included window adaptively, and tracking the multiple objects utilizing the matching results and Kalman filter. The detailed processing steps of the proposed object tracking system are as follows:

- Step 1: make a difference between the gray-level images of the reference frame and current frame and binarize the difference image.
- Step 2: obtain the multi-lateral histogram of the binarized image and process a smoothing for the multi-lateral histogram.
- Step 3: Determine the area that the object exists and detect the object in the area.
- Step 4: set up the search window to be slightly larger than the object window based on the object window that was controlled adaptively in the previous frame.
- Step 5: extract the invariant features of the objects using SIFT algorithm and then store features to the current feature buffer.
- Step 6: compute the matching rate of the feature values with location information between the previous and current feature vectors.
- Step 7: check whether the objects participate or not to the location matching. If yes, they are considered as the existing objects and go to next step 8, or not, they are guessed as a new objects or disappearing objects and jump to Step 9.
- Step 8: discriminate whether they are location-matched or not. If matched, the feature vector is stored to the new previous feature buffer that needs to match with the current feature buffer of the consecutive next frame, or not, they are rejected. And the object window is adjusted adaptively according to the size of the moving object.
- Step 9: the three representatives that the matching score is most high among the matched keypoints are contribute to the window size adaption and location

information of these representatives is applied as the inputs of Kalman filter.

- Step 10: In step 7, if the objects don't participate to the location matching, they are either appeared newly or no longer appeared, that is exit. If newly appeared, their features are stored the new previous feature buffer, or not, they are rejected.
- Step 11: repeat from step 4 to step 10 about consecutive next current frame image. These are repeated sequentially and the object tracking would be continued.

Figure 1 show the flow chart of the proposed object tracking system.

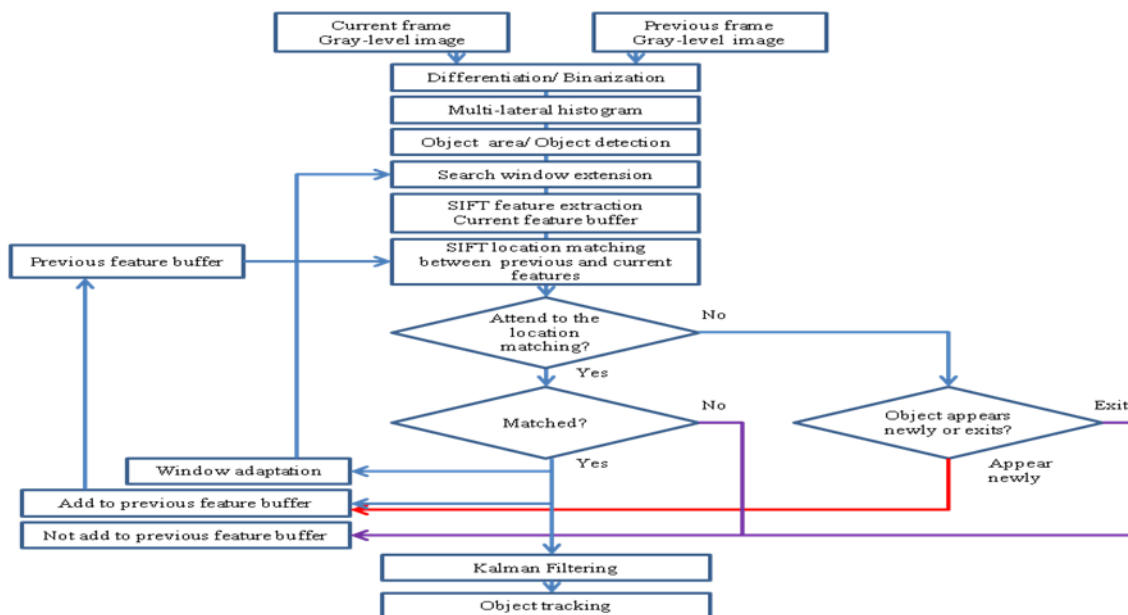


Figure1. Flow chart of the proposed object tracking

2.1 Object detection

Traditional lateral histogram is a statistic that integrated the brightness values of the pixels for the same row and column in the horizontal and the vertical axis directions of a gray-level image and it has been used to detect the existing area and the shape of a specific object.

In case that several objects appear in a frame image, a problem is caused that multiple objects are recognized as a single object because the histograms of these objects are overlapped in the direction they are standing. But, if each object regions are divided first from the horizontal axis histogram and then the vertical axis histogram is calculated about the individual horizontal regions, multiple objects areas could be obtained easily. This method is a simple extended one from the traditional lateral histogram and it is called multi-lateral histogram.

However even using this multi-lateral histogram it is actually not possible to discriminate the objects on conditions that multiple objects are close to each other or overlapped. Because these could be solved by matching their invariant features that be able to obtain by using SIFT algorithm, this simple multi-lateral histogram technique is very useful and concrete to detect multiple objects.

The process to detect the multiple objects areas through the multi-lateral histogram is as in the following:

- Step 1: from the horizontal histogram, process smooth filtering over the histogram and then obtain a threshold using the average of the total histogram values.

- Step 2: determine the region that the higher value than the threshold exist continuously as the object region and obtain the vertical histogram corresponding to the individual horizontal object regions and process smooth filtering
- Step 3: As the object area using the two object regions that crossing each other.

Figure 2 shows the object detection results through the two steps.



Fig. 2. Results of object detection by lateral histogram.

2.2 Invariant feature extraction and matching

Lowe's SIFT algorithm has been used widely in fields of object detection and recognition and has largely four steps as follows: Step 1: Scale space extrema detection: obtain the object's keypoints that invariant to size and orientations applying the DoG(Difference of Gaussians) function. Step 2: Keypoint localization: Select the keypoints that are stable about brightness and locations. Step 3: Orientation Assignment: assign the orientation information according to the brightness variation. Step 4: Keypoint description: describes the invariant feature information that was calculated from the brightness variation size and directions of the neighbor keypoints. Equation (1) presents the keypoint descriptor, that is the feature matrix, that includes the features extracted from SIFT.

$$[im, des, loc] = \text{SIFT}(\text{image}) \quad (1)$$

Where, im has the pixel values of the test image and des presents the matrix of the descriptor matrix, and loc has location, size, and orientation values of the all extracted keypoints. Figure 4 shows the invariant feature keypoints of the multiple objects extracted from SIFT. Upper line shows the original objects areas and the keypoint extraction results for the object area includes the background and lower line shows the results that are concentrated to the inside area of the objects. Here, when the computing is concentrated to the inside of the object SIFT generates the more and the reliable points comparing with the case of the object area. Traditional lateral histogram is a statistic that integrated the brightness values of the pixels for the same row and column in the horizontal and the vertical axis directions of a gray-level image and it has been used to detect the existing area and the shape of a specific object [6].

After the keypoints extraction of the multiple objects using SIFT algorithm, it is needed to compare the corresponding keypoints between the previous keypoints and the current ones for every objects using the loc information in equation (1) and it is called the keypoint location matching. The distance between two corresponding keypoints is calculated by their inner product as equation (2).

$$d_{ij} = \cos^{-1}(des_{ri} \cdot desc_j) \quad (2)$$

Where, des_{ri} and $desc_j$ are i th descriptor in the previous frame and j th descriptor in the current frame. The distance ratio that to be a reference of matching or mismatching is computed by equation (3).

$$\text{dist ratio} = \frac{\text{the colsest distance}}{\text{the second colsest distance}} \quad (3)$$

If this dist-ratio is smaller than the adaptive value 0.8 presented by Lowe's experiment it is considered as matching, if not, it is considered as mismatching. Figure 3 shows the results of the keypoint matching for a walking man and in

view of the matching line, we can find that there were a lot of mismatched keypoints nevertheless these are actually discriminated as the matched.

These mismatched keypoints are classified using the location information. We analyzed the reason of the mismatching and established a matching reference that if the distance between the two corresponding keypoints is smaller than the allowable distance error it is determined as the matching case, if not, it is determined as the mismatching case.

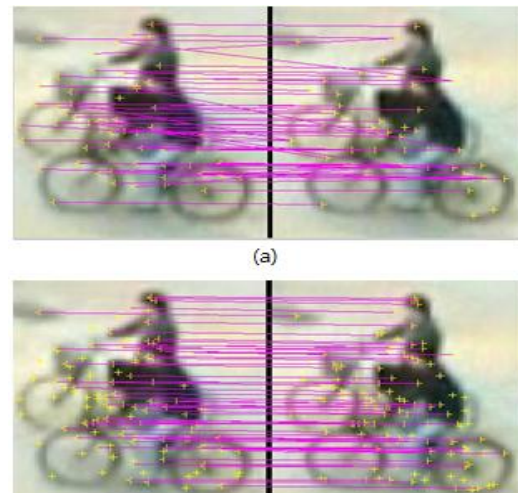


Fig. 3. Examples of keypoint mismatching and effective matching.

Because this matching reference uses the location distance information it is called as the location matching. By utilizing this reference the more exact matching is possible and ultimately it would be contributed to advance the more robust multiple objects tracking

2.3 Tracking window adaption

In multiple objects tracking, the objects are changed into various sizes and directions in process of time and therefore the window sizes for the keypoint extraction and matching are required to be controlled according to the object size adaptively. We propose a new method and it is a method that finds the extended rate of the changed size from the representative three matched keypoints of the two frame objects to be compared for matching. Figure 6 shows the geometrical figure of the proposed window adaption method.

Where, W and H are the width and the height of the current object area and $W/4$ and $H/4$ are the extended value for the moved object search and x and y are the locations of the starting point of the extended search, and dx and dy are offsets from the starting point to the starting point of the object area in the previous frame and the new object area in the current frame.

The processing steps are for tracking window adaption as follows:

Step 1: Select three location matched keypoints in the object area of the previous frame.

$$(x + d_{x11}, y + d_{y11}), (x + d_{x12}, y + d_{y12}), (x + d_{x13}, y + d_{y13}) \quad (4)$$

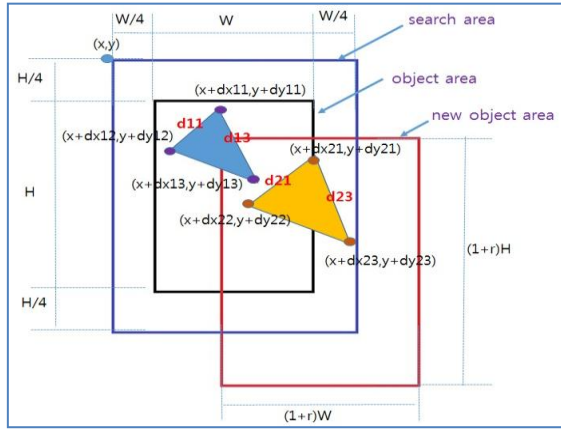


Fig. 4. Tracking window adaptation

Step 2: For the two pairs of keypoints, calculate the distance

$$d_{12} = \sqrt{(d_{x11} - d_{x12})^2 + (d_{y11} - d_{y12})^2} \quad (5)$$

$$d_{13} = \sqrt{(d_{x11} - d_{x13})^2 + (d_{y11} - d_{y13})^2}$$

Step 3: In the previous frame, select the search area by extending the object area by $W/4$ and $H/4$ in directions of the horizontal and vertical axis.

Step 4: In the current frame, select the search area that is identical location and size of the search area in the previous frame

Step 5: In the current search area, select three keypoints that is matched with the three keypoints selected in step 4.

Step 6: For the two pairs of keypoints in step 5, calculate the distance (Euclidean distance)

$$d_{21} = \sqrt{(d_{x21} - d_{x22})^2 + (d_{y21} - d_{y22})^2} \quad (6)$$

$$d_{23} = \sqrt{(d_{x21} - d_{x23})^2 + (d_{y21} - d_{y23})^2}$$

Step 7: Calculate the extended object scaling rates between the previous and the current frames, r_{12} and r_{13} and then calculate the average scaling rate r .

$$r_{12} = \frac{d_{21}}{d_{11}}, r_{13} = \frac{d_{23}}{d_{13}}, r = \frac{r_{12} + r_{13}}{2} \quad (7)$$

Step 8: Based on this average scaling rate, a new object window that includes the object in the current frame. The position of the new starting point is computed like next equations.

$$x \text{ position: } x + d_{x21} - r d_{x11} \quad (8)$$

$$y \text{ position: } y + d_{y21} - r d_{y11}$$

Step 9: Set the new object area that was controlled adaptively in the current frame. It has the size of $1+r$ W and $1+r$ H in the horizontal and vertical directions from the starting point.

Step 10: The windows of all objects are controlled adaptively using steps from 1 to 9.

These adaptive windows become the basis of tracking the multiple objects that appear in the next frame.

2.4 Object tracking

2.4.1 Feature matching using SIFT keypoints

Using the object feature matching that considers the location information the tracking procedure of the multiple objects is as follows.

Step 1: Store the keypoint feature information extracted in the previous frame.

Step 2: Test the matching rate between the feature data stored in the previous feature buffer and the current feature data extracted with identical size and location to the adaptive window selected in the previous frame.

matched= yes, if matched points ≥ 3

no, if matched points < 3 (9)

Here, we adopted 3 points as the threshold because the adaptive window is determined using only three keypoints.

Next two cases are the references that objects are presented newly in the current frame or that they are disappeared.

Case 1: Newly presented case, any object area with features that doesn't belong to the previous feature. This object feature is added to the new previous buffer for next matching and tracking.

Case 2: Disappeared case, this is a case that the comparable features don't exist in the current feature buffer. In multiple objects tracking, the objects are changed into

various sizes and directions in process of time and therefore the window

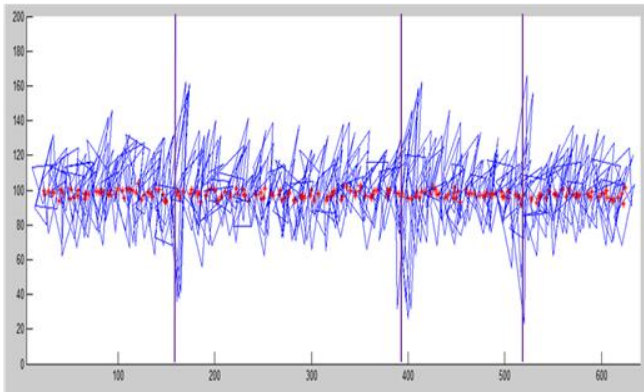


Fig. 5. Tracks by the three representatives and Kalman filter.

For real-time tracking, we used three representative keypoints that have the most matching rate among the matched keypoints and Figure 5 show the results of tracking using the representative keypoints for a series of 211 consecutive frames. Here, the blue triangles are the connectives of three representative keypoints and the red asterisk points are tracks by Kalman filter. At the 47th frame a severe fault track is getting generated because the object area is relatively enlarged due to the unstable distance extension and ultimately it may generate a tracking failure. So in this paper, for resolving this problem we composed the location information of the above three representative keypoints and the estimation characteristics of the Kalman filter.

2.4.2 Object tracking using Kalman filter

The Kalman filter, also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing noise and other inaccuracies, and produces estimates of unknown variables that tend to be more precise than those based on a single measurement alone. More formally, the Kalman filter operates recursively on streams of noisy input data to produce a statistically optimal estimate of the underlying system state.

The Kalman filter has numerous applications such as tracking, guidance, navigation and control of vehicles. This algorithm works in a two-step process. In the prediction step, the Kalman filter produces estimates of the current state variables, along with their uncertainties.

Once the outcome of the next measurement (necessarily corrupted with some amount of error, including random noise) is observed, these estimates are updated using a weighted average, with more weight being given to estimates with higher certainty. Because of the algorithm's recursive nature, it can run in real time using only the present input measurements

and the previously calculated state and its uncertainty matrix; no additional past information is required.

Time Update (prediction)	Measurement Update (correction)
$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k$	$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$
$P_k^- = AP_{k-1}A^T + Q$	$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$
	$P_k = (I - K_k H)P_k^-$

In this paper, we composed the location information of the above three representative keypoints and the estimation characteristics of the Kalman filter.

The three representative keypoints that selected from the matching step and these are presented as inputs to the Kalman filter. Through the estimation characteristics of the Kalman filter, the track failure due to the unstable distance extension will be stop and the tracking is maintained from next frame continuously.

In figure 5, we can find the tracking result by combining the Kalman filter and figure 6 shows the results of object tracking for three videos including a string of small objects.



Fig. 6. Results of object tracking using SIFT and Kalman Filter.

3 Experiments and results

We compared how do the number of keypoints generate in two cases of the object-included window area and the only object inside and table 1 show the results of two examples for one man and two bicycle men.

TABLE I. THE EXTRACTED KEYPOINTS NUMBER FOR TWO OBJECTS OF ONE MAN AND TWO BICYCLE MEN

	ONE MAN	TWO BICYCLE MEN
OBJECT WINDOW AREA	17	18
OBJECT INSIDE	24	73

In table 1, the number of the extracted keypoints in case of the object inside is much more than ones in case of the object window area. This means that concentrating to the inside of the object that doesn't include the reference background is more effective in the matching and tracking.

Next, we determined what is the most fare value of the distance for discriminating the location matched or not in the keypoint and the distance in here means the pixel value. Table 2 presents the numbers of the matched and mismatched keypoints according to the distance between the two comparable keypoints in the location matching for the one man case.

TABLE II. THE EXTRACTED KEYPOINTS NUMBER FOR TWO OBJECTS OF ONE MAN AND TWO BICYCLE MEN

Distance	1	2	3	4	5	6	7	8	9
Matched	1	3	16	19	20	21	23	25	25
Mismatched	0	0	0	0	0	0	1	1	1

Based on the results in table 2, we determined the fare distance range for the location mating and in this experiments we used 3 or 5. Lastly, using the proposed multiple objects tracking system the tracking performance is tested for three videos.

These videos include various cases that multiple objects close to each other or are overlapped in a variety types of movements and these videos were utilized by the sample videos with the frames more than 300 frames. Table 3 shows the tracking results for the three sample videos.

TABLE III. THE TRACKING RESULTS OF THE THREE VIDEOS BY USING ONLY SIFT WITHOUT THE KALMAN FILTER

	Success rate	Failure rare
Sample 1	92	8
Sample 2	89	11
Sample 3	87	13

In table 3, the tracking rate was about 89.3 on average and especially, in case of sample 3, even if this video has much complicated movements such as closing, overlapping, crossing, and occluding, relatively high tracking rate is generated.

TABLE IV. THE TRACKING RESULTS OF THE THREE VIDEOS BY COMPOSING SIFT AND KALMAN FILTER

	Success rate	Failure rare
Sample 1	99	1
Sample 2	100	0
Sample 3	97	3

Table 4 shows that the object tracking performance can be advanced greatly by composing the Kalman filtering with three representative ones that has most high matching rate among the matched keypoints as inputs of the filter.

4 Conclusions

A robust object tracking system was proposed that performs well in the complicated movements of multiple objects using the multi-lateral histogram, the SIFT algorithm, the location matching, the adaptive windowing, and Kalman filter. By combining Kalman filter and the representative keypoints of SIFT, an advanced object tracking system could be implemented. Based on the several experimental results the proposed system has a robust characteristic to track multiple objects. In future, it is needed to advance the processing speed and it will be considered to implement the system using the parallel processing technique for the real-time tracking.

5 References

- [1] D. G. Lowe, "Object Recognition form Local Scale-Invariant Features," Proc. of the International C onference on Computer Vision, pp. 1150-1157, 1999
- [2] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, pp. 1615-1630, 2005
- [3] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," Proc. of the Ninth European Conference on Computer Vision, pp.404-417, 2006.
- [4] H. Seokwun, "Multiple objects tracking with location matching and adaptive windowing based SIFT algorithm," International journal of Advanced Computer Technology, pp.427-432, 2013.
- [5] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering Vol. 82, No. 1, pp. 35-45, 1960.
- [6] R. Davis, "Lateral histogram for efficient object location: Speed versus ambiguity," Pattern Recognition Letters, Vol. 6, No. 3, pp.189-198, 1987

Subjective Video Streaming Quality Evaluation in 3G Cellular Networks

Fahad Al Qurashi, Hamad Almohamedh, Ivica Kostanic

Department of Electrical and Computer Engineering

Florida Institute of Technology

Melbourne, Florida, USA

falqurashi2008@my.fit.edu, halmoham@my.fit.edu, kostanic@fit.edu

Abstract - A novel method for subjective QoS evaluation of streaming video is developed. The method, named Mobile Video Quality Prediction (MVQP), relies on a non reference QoE measuring tool. This is a hybrid between subjective and objective measurement. To validate this approach as a part of the MVQP project, a live video streaming platform is designed. The platform includes different video sequences (high and low motions). The results presented in this paper are based on subset of eight videos selected from the MVQP video quality database. A commercial, Sprint PCS 3G mobile networks was used for test evaluated. fifty subjects evaluated video quality using smart phones based on ITU recommendation. Mean Opinion Score (MOS) was compared with packet loss at the end of this study.

Keywords: MVQP, Subjective test, 3G cellular networks

1 Introduction

The data traffic over cellular networks owes a great deal of its rise to the video-streaming applications. These applications cause a fluctuation on the data rates, which in turn influence the quality of streaming [1]. Ever since the world has been struck with massive usage of mobile based technologies, various users are using their mobile phones to watch videos. This has led to a revolutionary increase in mobile-video streaming amongst a large number of users [2]. The video streaming quality assessment is not an easy task. This assessment faces various challenges like limited computation powers of the mobile device [3].

Traditional Quality of Service (QoS) does not offer adequate means for assessing real-time video streaming applications. This is because an important attribute of user satisfaction cannot be assessed under such a method [4]. Multimedia techniques for quality measurement are based on the Quality of Service (QoS). However, it is deemed that these metrics of the traditional QoS do not sufficiently measure the actual quality of the multimedia applications; and therefore, changes will have to be incorporated in the future. One may say that the QoS metrics are purely network-based and as a result they cannot measure factors as perception and sensation of the user.

For that reason, the QoE is regarded as a more effective method for assessing the quality perceived by the end-users [5]. QoE is a measurement of the quality of satisfaction that is derived from a communication system [6]. QoE denotes how users perceive the quality of such running applications and the user perceived experience of the service provided by the network provider.

2 MVQP Project (3G)

Estimation of the video streaming quality in 3G Cellular Networks over smart phone device is an issue that has not been satisfactory addressed. At the moment not many solutions have been proposed. Partially, this may be attributed to computational, size and other inherent limitations of the mobile devices. The main goal of MVQP is to develop a new method for the prediction of video streaming quality through mobile networks. The method will be referred to as Mobile Video Quality Prediction (MVQP).

Subjective evaluation is the best method for the assessment of the video streaming quality. In subjective assessment, a human subject evaluates if the quality of the video is good, average, or poor. MVQP will use the subjective assessments to calibrate mapping between a set of objective engineering measurements and subjective QoE.

MVQP will be able to recognize the quality of the streaming video for network providers as an estimation for the quality of streamed video in mobile connections utilized by mobile devices. MVQP Technology will completed in two phases as shows in Figure 1.

Phase one contain development of MVQP raw video database and a set of factors that affecting video quality. MVQP aims not just to propose but also deliver a new scheme to predict the quality of the video streaming using the 3G Cellular Networks.

In phase two MVQP tools will be able to recognize the quality of the streaming video for network providers as an estimation for the quality of streamed video in 3G connections utilized by the smart mobile devices. After phase one, the quality affecting video streaming factors will be collected from this study. These factors will be used to

train and validate the MVQP algorithms. The ultimate goal is development of the automatic prediction of the user experience which is sometimes referred to as the E-MOS.

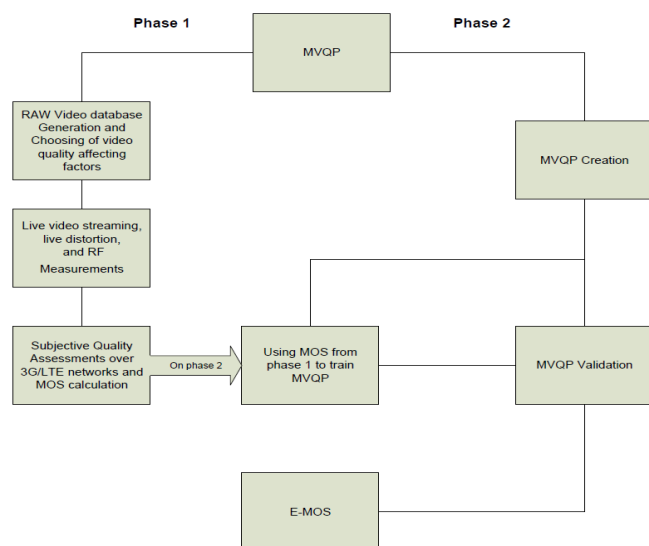


Figure 1. MVQP chart

3 Quality affecting factors

For this study, a set of factors that affect the QoE are selected. These set contains the packet loss with some Radio Frequency (RF) factors. It is assumed that the streaming video is delivered over UDP and hence, the delays usually found in the TCP are not relevant.

3.1 Packet loss

Packet loss is defined as a rate at which transmitted packets do not reach their destination. Packet loss is the most important objective factor affecting the quality of receiving videos [8].

3.2 Received signal code power (RSCP)

Received signal code power (RSCP) is the measure of power at the receiver. It is attributed to a specific physical communication channel. It signifies the strength of the signal used for data delivery [9].

3.3 Interference metric (Ec/Io)

(Ec) is the received pilot energy, (Io) is the total power spectral density or alternately the total received energy. Pilot quality is defined as the Ec to Io ratio that is expressed in dB [10]. This is a fundamental metric of the signal quality.

4 Live video streaming, live distortion, and RF measurements methodologies

In this study, eight raw videos from MVQP database [11] are selected. Different videos were selected to keep the test bed as diverse as it could possibly be and to ensure the results are obtained against a representative set of videos. The 4K raw video is down sampled to proper resolution for 3G bandwidth which is 480x320 by using H.264 codec in 30fps. The frame quality of a video is usually indicated by codec and bit-rate [4].

An efficient coding and reduction of bit-rates is recently developed in the video compression standard H.264. The H.264 shows significant improvements over the older standards of H.263 and MPEG-4 [7]. As an example, H.264 has some built in features that reduce the errors of coding such as providing small sizes of blocks and filters for de-blocking covered by [7].

The factors kept under consideration during the experiment are the Packet loss, RSCP, and (Ec/Io). A "CDMA2000/ EVDO-Rev A" on Sprint PCS provider with a bandwidth Up to 3.1 Mbps was used to stream the videos and to measure the impact of the above mentioned distortion factors. The results revealed a relationship between the factors, which is shown in part five of this paper.

4.1 Videos source

The lists below are a short description for eight of raw videos from MVQP database [11], and the snapshot of each of them are shown in Figure 2.

4.1.1 Soccer game (sg)

Shot on a campus on a sunny afternoon. Players are showing diverse contrasts and colors along with complex motions. The camera is tracking the players both sides horizontally.

4.1.2 Lawn service (ls)

Shot on campus on a sunny morning. A man is providing lawn services by making use of a lawn machine. The camera tracks the machine from left to right.

4.1.3 Pedestrian (pe)

Shot on campus on a sunny morning. Some students are entering while others are leaving. The camera was fixed.

4.1.4 Garden (ga)

Shot in a garden on a cloudy morning. There are light color contrasts and slow motion of tree leaves. The camera tilts the trees from bottom to top with a reflection of the cloudy sky.

4.1.5 Turtle (tu)

Shot at a lake on a cloudy morning. The slow movement of turtle within the water has presented a fascinating scene. The camera was slowly tracking the turtle.

4.1.6 Large building (lb)

Shot on campus on a sunny morning. Many small leaves are visible, moving slowly in different directions. The camera was moving from the bottom of the building towards the blue sky diagonally.

4.1.7 Bridge (br)

Shot across the bridge in the city of Melbourne on a cloudy afternoon. Various cars are moving on the bridge while water waves are moving slowly downstream. The camera was fixed.

4.1.8 Basketball (ba)

Shot in a park basketball field on a sunny afternoon. Children are getting trained for basketball and a complex motion is being depicted through their movements. The camera was fixed.



Figure 2. RAW Videos snapshots [11]

4.2 Measurements hardware

The computer server that was used in this study is an Intel based PC with 1TB hard drive and 8GB of RAM. The operating system is Linux. Internet speed average is 256 Mbps for uploading and downloading. The computer client is a ThinkPad laptop with Intel processor and has storage of 240 GB SSD with Windows 7 OS. It is set up to receive mobile video streaming from the server via Samsung Galaxy S4 while connected with a USB cable.

FEMPEG, an open source tool installed in both server and client, does live streaming video and stores the video files from the server.

4.3 Measurements applications

MVQP measurement application for collection of the radio frequency (RF) parameters was developed over android platform. Using Samsung mobile device the measurements are collected once per second. The measurement application collects Ec/Io and RSCP.

4.4 Subjective methodology

In subjective quality testing, a set of sequences of the processed video is set for evaluation before the human observers. There are several methods identified by the research community for testing. In case of videos, standards methods are defined in order to conduct subjective tests for television broadcasting standards [13] [14].

In the situation of a subjective quality evaluation for mobile-videos, a lot of efforts are required in order to examine the environment of the test, the display's set-up, test material assessment and the regulation of adequate viewing distance for the viewer [15].

The Absolute Category Rating (ACR) method is used in the quality tests. This method is a category judgment in which one test condition is presented only once to the viewers. This method is also known as Single Stimulus 2 Method. On the basis of category scale test sequences are rated separately. After each presentation viewers are asked to assess sequence's quality.

Figure 3 describes the time pattern for the incentive presentation. In the voting mechanism voting time should be less than or equal to the ten seconds. Time for the presentation can be decreased or increased on the basis of test material's content [13]

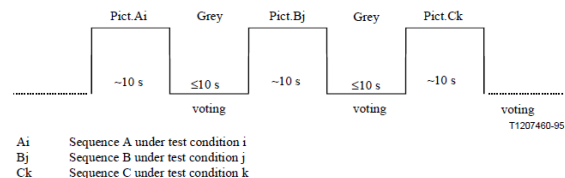


Figure 3: Stimulus presentation in the ACR method [13]

In this scheme after viewing or listening presentation subjects are asked to rate the presentation's quality. This stage is known as voting time. For rating most commonly used scale is five-level scale [16].

In this subjective test, fifty subjects participated with a real mobile devices by using our MVQP android subjective

test application. The application collected the subjective rating after showing them random sequence number of distorted videos. The sequence videos have been distorted in the live measurement streaming step.

Five Samsung Galaxy mobiles were used in this subjective test. The university lab was used with a five partition tables for the 160 rounds of test sequence. The subjects were divided into groups. Each subject rated 16 of a randomly distorted video collection sequence. A short presentation was given before each session. Each session was sent to the MVQP server immediately after each rating was stored in rating database.

5 Results

The results shown in Figs 4-11 are obtained using eight representative videos from the MVQP video database. Some of them are fast motion and the others are slow motion. Sprint PCS mobile network is used for our live video streaming experiment.

It is evident from the graphs of that the packet loss is higher if the signal (RSCP) is either high or low with the presence of interference (Ec/Io). Also, the MOS confirmed that if the packet loss is high then the MOS drops down.

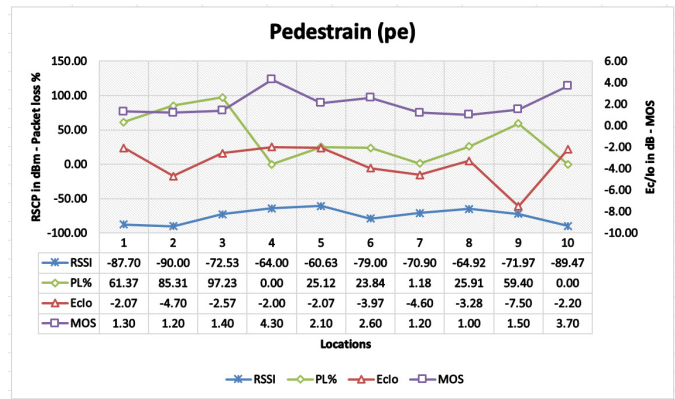


Figure 6. Pedestrian (pe) video streaming analysis

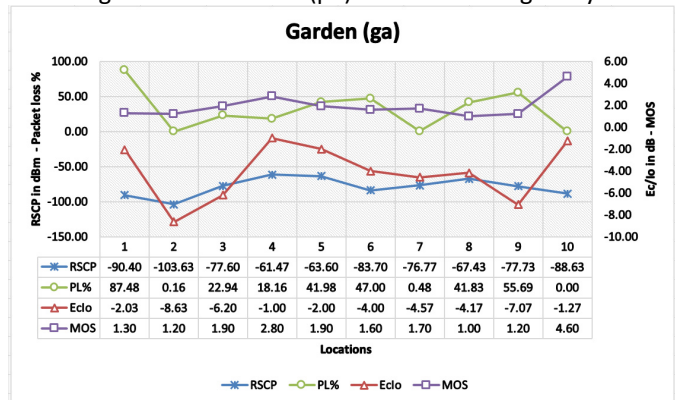


Figure 7. Garden (ga) video streaming analysis

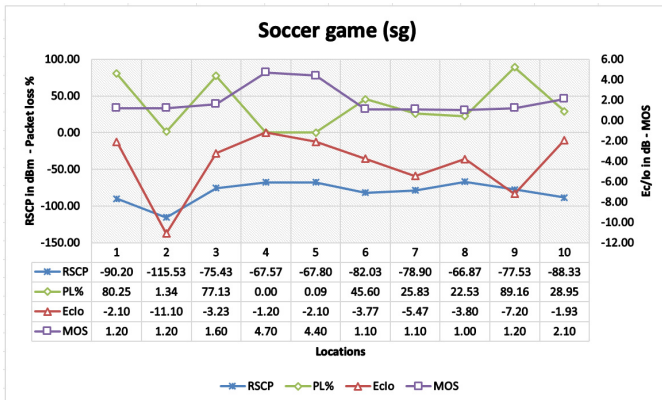


Figure 4. Soccer game (sg) video streaming analysis

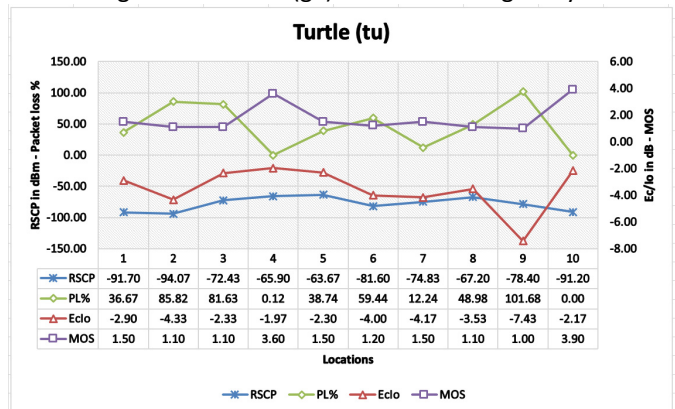


Figure 8. Turtle (tu) video streaming analysis

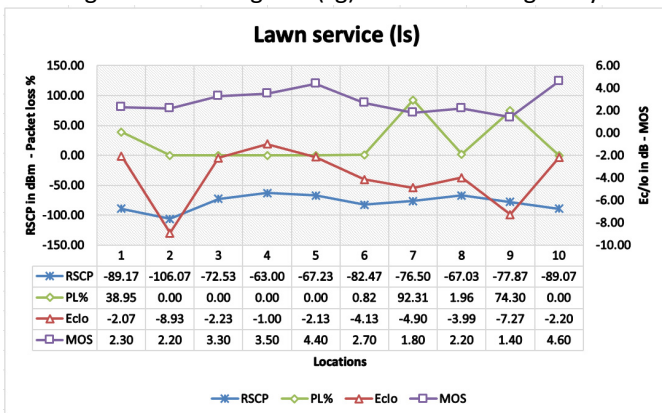


Figure 5. Lawn service (ls) video streaming analysis

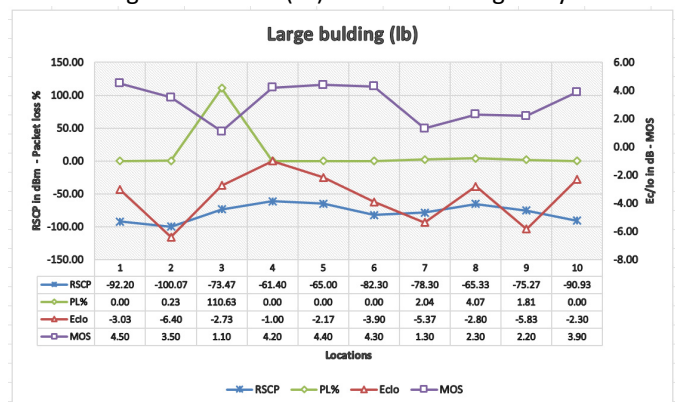


Figure 9. Large building (lb) video streaming analysis

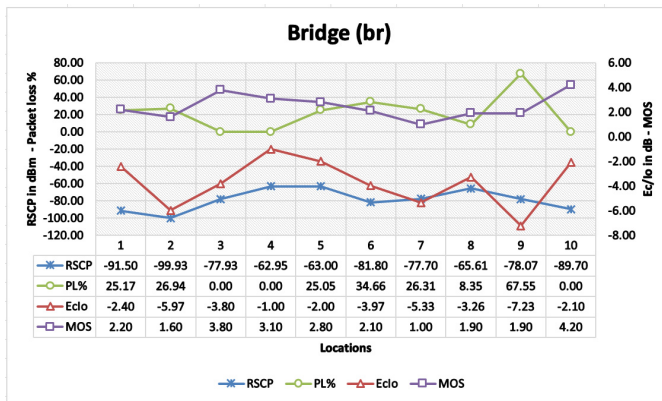


Figure 10. Bridge (br) video streaming analysis

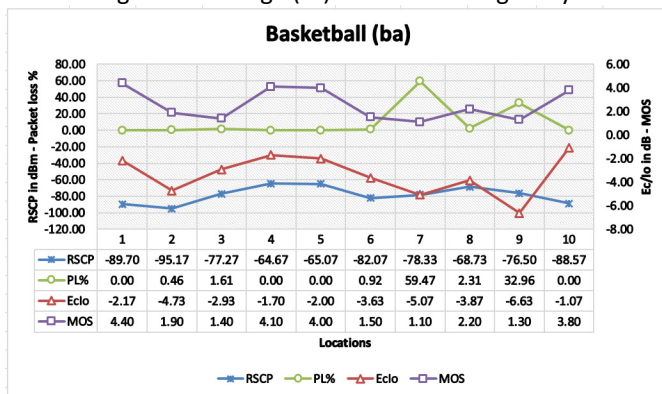


Figure 11. Basketball (ba) video streaming analysis

6 Conclusion and future works

This study evaluated the third stage of MVQP project (phase one). The study was backed up by the experimental results which showed a relationship between the considered distortion factors. In the future MVQP (phase two), a QoE tool will be built that can predict the video quality from the results that were collected from phase one. In [12] and [17] we extended this study with another experiments that contained the in depth analysis of the experiments for LTE video quality streaming measurements.

7 References

- [1] Ayaskant Rath, Sanjay Goyal, and Shivendra Panwar; "Streamloading: Low Cost High Quality Video Streaming for Mobile Users"; Department of Electrical and Computer Engineering, Polytechnic Institute of New York University
- [2] Wei Song, Dian Tjondronegoro, Michael Docherty; "Quality delivery of mobile video: in-depth understanding of user requirements"; OzCHI '11: Proceedings of the 23rd Australian Computer-Human Interaction Conference; November 2011
- [3] An (Jack) Chan, Amit Pande, Eilwoo Baik, Prasant Mohapatra; "Temporal Quality Assessment for Mobile Videos"; Mobicom '12: Proceedings of the 18th annual

international conference on Mobile computing and networking; August 2012

- [4] Ghareeb, M.; Viho, "Hybrid QoE Assessment Is Well-Suited for Multiple Description Coding Video Streaming in Overlay Networks "C. Communication Networks and Services Research Conference (CNSR)", 2010 Eighth Annual Digital Object Identifier: 10.1109/CNSR.2010.15 Publication Year: 2010, Page(s): 327 - 333

- [5] Aguiar, E.; Riker, A.; Mu, M.; Zeadally, S.; "Real-time QoE prediction for multimedia applications in Wireless Mesh Networks"; Cerqueira, E.; Abelem, A. Consumer Communications and Networking Conference (CCNC), 2012 IEEE Digital Object Identifier: 10.1109/CCNC.2012.6181017 Publication Year: 2012, Page(s): 592 - 596

- [6] Meral Shirazipour, Gregory Charlot, Geoffrey Lefebvre, Suresh Krishnan, Samuel Pierre ; "ConEx based QoE feedback to enhance QoE"; CSWS '12: Proceedings of the 2012 ACM workshop on Capacity sharing; December 2012

- [7] Satu Jumisko-Pyykkö ; "Evaluation of subjective video quality of mobile devices"; MULTIMEDIA '05 Proceedings of the 13th annual ACM international conference on Multimedia Pages 535 - 538

- [8] Wei Song, Dian Tjondronegoro, Michael Docherty; "Quality delivery of mobile video: in-depth understanding of user requirements"; OzCHI '11: Proceedings of the 23rd Australian Computer-Human Interaction Conference; November 2011

- [9] An improved mobile location approach based on RSCP difference Zhenqiang Wang ; Yi-Sheng Zhu ; Nan Liu "Electronic and Mechanical Engineering and Information Technology (EMEIT)", 2011 International Conference on Volume: 6 Digital Object Identifier: 10.1109/EMEIT.2011.6023675 Publication Year: 2011 , Page(s): 2765 – 2768

- [10] Rockstar Bidco Lp; "System and Method for Ec/Io Access Screening in a CDMA Network" in Patent Application Approval Process Politics & Government Week (Jul 25, 2013): 2666.

- [11] Fahad Al Qurashi, Hamad Almohamedh, and Ivica Kostanic "RAW Video Database of Mobile Video Quality Prediction (MVQP)"; The 2014 International Conference on Image Processing, Computer Vision, and Pattern Recognition (in-press), 2014.

- [12] Hamad Almohamedh, Fahad Al Qurashi, and Ivica Kostanic "Mobile Videos Quality Measurements over Long Term Evolution (LTE) Network"; The 2014 International Conference on Image Processing, Computer Vision, and Pattern Recognition (in-press), 2014.
- [13] "ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications, 2000."
- [14] "ITU-R Recommendation BT.500-11. Methodology for the subjective assessment of the quality of television pictures."
- [15] Hands, D.S.; Huynh-Thu, Q. ; Rix, A.W. ; Davis, A.G. ; "Objective perceptual quality measurement of 3G video services "; Voelcker, R.M. 3G Mobile Communication Technologies, 2004. 3G 2004. Fifth IEE International Conference on Digital Object Identifier": 0.1049/cp: 20040712 Publication Year: 2004, Page(s): 437 – 441
- [16] Mohamed, S.; Rubino, G. "A study of real-time packet video quality using random neural networks "; Circuits and Systems for Video Technology, IEEE Transactions on Volume: 12, Issue: 12 Digital Object Identifier: 10.1109/TCSVT.2002.806808 Publication Year: 2002, Page(s): 1071 - 1083 Cited by: Papers (61) | Patents (1)
- [17] Hamad Almohamedh, Fahad Al Qurashi, and Ivica Kostanic "Subjective Assessment of Mobile Videos Quality for Long Term Evolution (LTE) Cellular Networks" (in-press), World Congress on Engineering and Computer Science 2014 (WCECS 2014).

On Sparse Representation and Visual Tracking

Jingya Wang¹ and Shahram Payandeh¹

¹School of Engineering Science, Simon Fraser University, Burnaby, British Columbia, Canada

Abstract—In recent years, sparse representation has been widely applied to areas such as machine learning, pattern recognition, computer vision and image processing. Sparse representation, in general, uses the fewest elements from a specified dictionary to acquire, represent and compress high-dimensional signals. This paper first presents an overview of sparse representation and its usage in various fields and then reviews how such representation is utilized in the field of visual tracking.

Keywords: sparse representation; visual tracking; dictionary learning.

1. Introduction

In the history of engineering, when modelling systems or processing signals, searching for a canonical model of certain process or representation of signals using a collection of simpler models has been challenging pursuits. Image as a form of signal, is analysed, compressed and de-noised by powerful tools such as Fourier transform, wavelet and principal component analysis (PCA). The goal of these methods is to find a set of bases to represent signals or images.

The aim of these transforms is often to reveal certain structures of a signal and to represent these structures in a compact and sparse representation [1]. Sparse representation defined as the most compact representation of a given signal which consists of a linear combination of the fewest signal bases which called atoms in an over-complete dictionary. As showed in Figure 1, a 1-D signal can be reconstructed by a linear combination of Fourier bases.

Before sparse representation, the objective has been in finding the fewest orthogonal bases which was taken as the best way to reconstruct the signal. Now the concept have changes. In stead of finding the fewest orthogonal bases we are try to build an over-complete dictionary where the vectors are not independent to each other. And the goal is solving for the sparsest coefficient vector to represent the signal. For example, we take Fourier bases as a dictionary and a 1-D signal x can be represented as a linear combination of these bases. Shown as figure 1, x equals to the bases in the dictionary multiply with a coefficient vector α which weighs the Fourier basis. If there is a signal that exactly identical with the signal x in the over complete dictionary, the coefficient vector α now is a sparse vector that full of zero except 1 non-zero element (Figure 2). A sparse representation can be obtained by a given over-complete

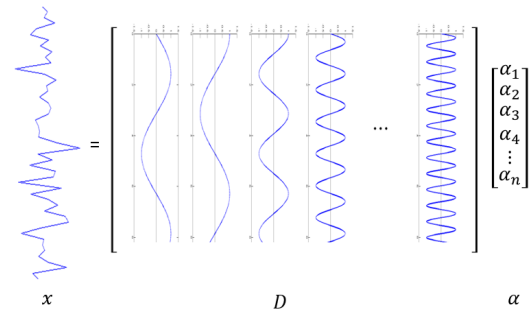


Fig. 1: Fourier transform of a given signal. 1-D signal x can be represented as a linear combination of Fourier basis.

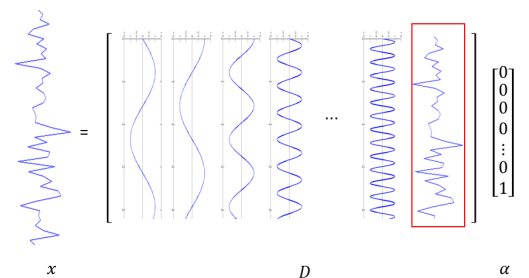


Fig. 2: Sparse representation with an over-complete dictionary. With an over-complete dictionary, the coefficient vector α now become sparse.

dictionary.

In 1996 a research done by Bruno Olshausen and David Field [2] have determined a possible coding strategy of our primary visual cortex. It was shown that there are millions of neurons in the cortex which are responding to different colors, shapes and textures exist within an image where only few of these neurons are activated for representing the image. These neurons construct an over-complete dictionary of the scene. The way the cortex represent images is called sparse representation which implies usage of fewest neurons or elements of the dictionary to reconstruct images. The notion of sparseness is also motivated in other areas such as natural language processing. In natural languages, to express the same meaning, a richer dictionary can help to construct a shorter sentence [3]. From these two examples, it can be observed that an image which is acquired based on millions of neurons is analogous to having a rich dictionary used in natural language processing.

At the heart of sparse representation lies a simple linear system of equations. A full-rank matrix $\mathbf{D} \in \mathbb{R}^{n \times m}$ which generates an under-determined system of linear equations expressed as $\mathbf{D} \alpha = \mathbf{x}$ which can have infinitely many solutions. We are interested in seeking its sparsest solution, i.e., the one with the fewest non-zero entries. The questions which remain are: Can such a solution ever be unique? If so, when? How can such a solution be found in reasonable time [1] ?

Early idea of sparse representation appeared in a pioneering work by Stephane Mallat and Zhifeng Zhang in 1993 [4], with the introduction of the notion of dictionaries which replaces the more traditional notion of wavelet transform. Their work put forward some of the core ideas that later became central in this field, such as a greedy pursuit technique that approximates a sparse solution to an under determined linear system of equations.

A second key contribution was made by Scott Shaobing Chen, David Donoho, and Michael Saunders in 1995 [5], who introduced another pursuit technique that uses the l_1 -norm for evaluating sparsity. They have shown that the quest for the sparsest solution could be tackled as a convex programming task, often leading to the proper solution [1]. Due to great advantage of sparse representation in image processing, it has been widely used on image restoration [6], inpainting [7], deblurring [8] and object tracking [9], [10]. The rest of this paper is organized as follows. Section II presents an overview of the sparse linear model and solution methods. Section III introduces sparse representation as it applied to object tracking. Section IV presents discussions and conclusions.

2. Sparse Representation

The core element of sparse representation is the linear system of equations $\mathbf{x} = \mathbf{D}\alpha$. Where vector $\mathbf{x} \in \mathbb{R}^m$, $\alpha \in \mathbb{R}^n$, $\mathbf{D} \in \mathbb{R}^{n \times m}$ and $n \ll m$, usually \mathbf{D} is full rank. The goal is solving α under known \mathbf{x} and \mathbf{D} . This problem is not difficult to solve, because of $n \ll m$, so this linear system of equations is undetermined. As a result this system having infinitely many solutions [11]. But if we want to seek for the sparsest solution for α implies in having the least non-zero element in α , implies $\|\alpha\|_0$ should be as small as possible. In other words, the problem is given an over-complete dictionary \mathbf{D} , how to find the least bases coefficient vector α to reconstruct the target vector \mathbf{x} .

$$\min \|\alpha\|_0 \text{ s.t. } \mathbf{x} = \mathbf{D}\alpha \quad (1)$$

In 2004, Donoho and Elad proved that if \mathbf{D} satisfied a given condition, then there exists a unique solution for equation (1) [12]. Although the uniqueness can be proven, the problem is still NP hard, and difficult even to approximate. To make this problem solvable, Terrence Tao and Candes proved that

if the solution α is sparse enough, the solution of equation (1) equals to the solution of l_1 minimization problem:

$$\min \|\alpha\|_1 \text{ s.t. } \mathbf{x} = \mathbf{D}\alpha \quad (2)$$

Actually l_1 minimization problem is convex, therefore the solution of equation (2) is unique [13]. By considering the effect of noise, equation (1) become

$$\min \|\alpha\|_0 \text{ s.t. } \|\mathbf{D}\alpha - \mathbf{x}\|_2 \leq \epsilon \quad (3)$$

which can have similar solution.

2.1 Sparsest Solution

In general, solving the sparsest solution is not easy especially given a large \mathbf{D} . To simplify the solution process, reduction in time and improve accuracy, a number of optimization algorithms based on equation (3) have been proposed in order to find the sparsest solution. The main class of methods are greedy procedures [4], homotopy[14], [15], soft-thresholding based methods[16], [17], re-weighted- l_2 methods[18], active-set methods [19]. Basically, they fall into two categories.

a) Sparsest solution via l^0 -minimization

These methods can be considered as directly optimization based on equation (3). The most common algorithm here is greedy algorithm, for example Matching Pursuit and Orthogonal Matching Pursuit. Given a fixed \mathbf{D} , matching Pursuit will sequentially chose the one atom that has the biggest inner product with the \mathbf{x} , then subtract the contribution due to that atom and repeat this process until the input signal \mathbf{x} satisfactorily decomposed.

b) Sparsest solution via l^1 -minimization

As mentioned in the above, we can replace l_0 minimization problem with a l_1 minimization problem which is convex and can be solved in polynomial time by standard linear programming method.

$$\min \|\alpha\|_1 \text{ s.t. } \|\mathbf{D}\alpha - \mathbf{x}\|_2 \leq \epsilon \quad (4)$$

methods based on soft-thresholding [20], [21] or Homotopy [22] can be used to solve for the sparsest solution.

2.2 Dictionary Learning

Given the general equation of the form $\mathbf{x} = \mathbf{D}\alpha$, where the structure of \mathbf{D} is unknown. The objective is how to reconstruct the structure of \mathbf{D} given a set of sample signals \mathbf{x} . This is a common condition in signal and image processing. How should we construct a proper over-complete dictionary \mathbf{D} which can reconstruct the training set with sparse representation and minimize the empirical or expected cost function? This process is defined as dictionary learning. It is not a joint optimization problem with respect to the dictionary \mathbf{D} and the coefficients α , which is not a jointly

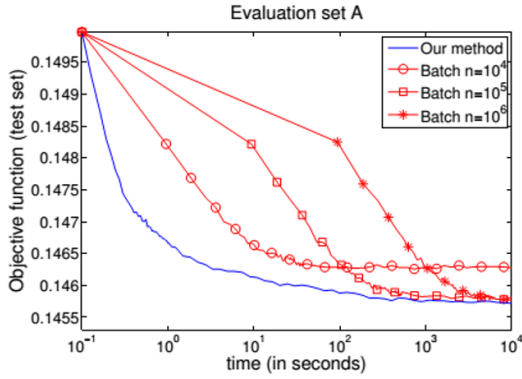


Fig. 3: Comparison between online and batch learning for various training set sizes [23]. The x axis indicates the running time and y axis shows the test error.

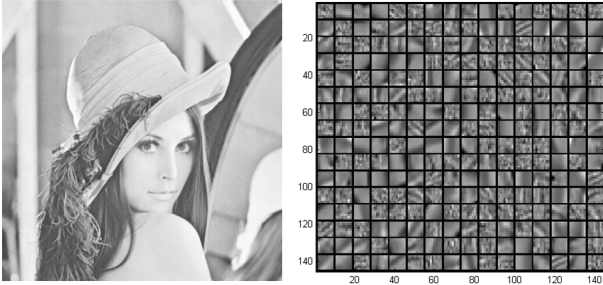


Fig. 4: Dictionary bases (right) learned from the left picture.

convex [23], but convex with respect to each of the two variables \mathbf{D} and α when the other one is fixed:

$$\min_{x_i, D} \sum \frac{1}{2} \|x_i - D\alpha_i\|_2^2 + \lambda \|y_i\|_1 \quad (5)$$

So a general method to solving this problem is to alternate between the two variables, minimizing over one while keeping the other one fixed. First, finding the coefficients α given the dictionary \mathbf{D} . Then, the dictionary is updated assuming known and fixed coefficients [24]. Repeating these two steps until the algorithm converges to some local minimum. There are three popular methods for dictionary learning, these are: efficient sparse coding algorithm [25], K-SVD [26] and online dictionary learning for sparse coding [23]. The first two methods are using batch learning which are computationally demanding. Compare to the first two methods, online dictionary learning is much faster by which process the signal one at a time or in mini-batches. Figure 3 compares the speed of online dictionary learning and batch learning. For example, the right side of figure 4 shows the over-complete dictionary that was learnt from the left side image based on online dictionary learning.

3. Object tracking via sparse representation

Tracking is an important component in many applications such as surveillance, human computer interaction, vehicle navigation, etc. [27]. Many challenges such as loss of information caused by projection of the 3D world on a 2D image, noise in images, complex object motion, non-rigid or articulated nature of objects, partial and full object occlusions, complex object shapes, scene illumination changes and real-time processing requirements still exist in order to design and develop a robust tracking algorithm for various given applications.

To address these challenges, early methods such as Mean Shift (MS) tracker [28], covariance (CV) tracker [29], and appearance adaptive particle filter (AAPF) tracker [30] were proposed. One of the pioneering work [31] applied sparse representation to face recognition. They proposed and built a redundant dictionary of a training subject taken under varying illumination conditions.

Object tracking works in similar ways [9]. Inspired by [32], the work presented in [9] was the first apply to sparse representation in tracking problem. To handle noises and occlusions, an error \mathbf{e} is added in the approximation of the sparse solution. So $\mathbf{x} = \mathbf{D}\alpha$ would be modified as

$$\mathbf{x} = \mathbf{D}\alpha + \mathbf{e} = \begin{bmatrix} \mathbf{D} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \alpha \\ \mathbf{e} \end{bmatrix}. \quad (6)$$

This paper builds a dictionary \mathbf{D} by stacking template image columns to form a 1D vector. In this paper, they use the gray value of each pixel of 10 template patches to form the dictionary (Fig.5). And the sample images \mathbf{x} is drawn from the video stream based on particle filter. Then l^1 -regularized least squares approach was adopted to find the sparsest solution in their vehicle tracking problem to locate the best tracking result. Compare to early methods, their approach demonstrates promising performance. A sparse solution of a good candidate image and a non-sparse solution of a bad candidate images are also shown in Figure 5. The drawbacks of this approach are extensive computational cost and the expressiveness of this model which is hard to handle view or pose changes makes this approach is not practical.

Instead of using normalized raw images to build the dictionary, [33] proposed a two stage sparse optimization. In the first stage, they utilize a dynamic group sparsity (DGS) and greedy algorithm to find and weigh the features that can discriminate the target and background. This first stage would project the raw image space into a feature space that are most discriminative in separating the target from background and also reduces the dictionary from high dimension to low dimension. The second stage solved the reconstruction coefficient vector α by optimized l^1 -minimization algorithm. The algorithm in [33] minimizes the target reconstruction error and maximizes difference between

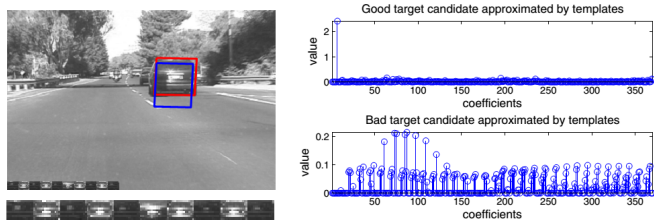


Fig. 5: The dictionary is constructed by 10 template shown in the bottom left corner of the top image. Top right: good target candidate approximated by template set. Bottom right: bad target candidate approximated by template set [9].

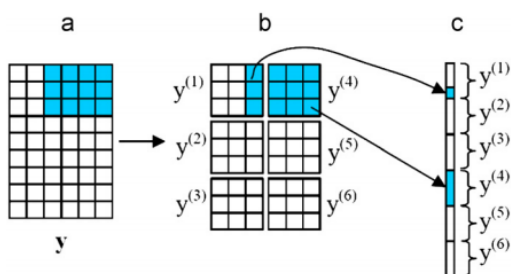


Fig. 6: a) Observed sample or template image (partial occlusion marked as blue). b) Divide the target and template into image blocks. c) Stack the blocks into 1D vector [10].

the target and background.

More recently, [34] explored an alternative appearance model formulation with sparse representation, which casts the tracking problem as finding a sparse representation of sub-image feature set template around the target. The tracking candidate is obtained by Kalman Filter. The tracking result is associated with the candidate holding the sparsest representation with the templates. This method is successful in the experiments. However, it is not clear if such method is able to track objects effectively when they undergo significant illumination and pose variations.

Another approach which called structured sparse representation is proposed by [10] is to divide the target and template into image blocks (Fig.6). Because of the partial occlusion often appears as a contiguous spatial distribution in the observed target sample, the occlusion can be stacked as a block sparse vector that has clustered non-zero entries. This approach of building the dictionary would be more robust to occlusion, rotation and pose changes.

Similar to structure sparse representation, [35] employed a multi-part sparse reconstruction code. The method is used on segments of the target instead of the whole target, and implemented by solving an l^1 -regularized least squares problem. The segment group with the smallest projection error will be taken as the tracking result. Compare to [9] which treat a target as a whole, [36] divides a target into

several parts and compare the distance between each part and the template, which act as good performance on tracking a target in random crowded scenes tracking.

To address tracking challenges such as appearance variation of object and its background, partial occlusion or deterioration in object images occurs, Kalman Filter are usually adopted to update the template space to keep the temporal consistency and adaptation to appearance variation.

4. Discussions and Conclusions

In this paper we have reviewed basic ideas and methods of sparse representation and its application in object tracking. The framework for all the tracking algorithms which adopts sparse representation are focused on two steps: 1) selection of good features to build the dictionary which also try to reduce its dimension; 2) the selected optimization algorithm for solving the sparsest solution. Sparse representation improved accuracy and reduced processing time in visual tracking. And it also points a new direction in visual tracking. With optimization such as structured sparse representation, sparse representation could obtain roust result of tracking object in extreme pose and illumination variation.

References

- [1] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.
- [2] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
- [3] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Access Online via Elsevier, 2008.
- [4] Stephane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.
- [5] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.
- [6] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *Image Processing, IEEE Transactions on*.
- [7] Mohamed-Jalal Fadili and Jean-Luc Starck. Em algorithm for sparse representation-based image inpainting. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–61. IEEE, 2005.
- [8] Weisheng Dong, D Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *Image Processing, IEEE Transactions on*, 20(7):1838–1857, 2011.
- [9] Xue Mei and Haibin Ling. Robust visual tracking using l_1 minimization. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1436–1443. IEEE, 2009.
- [10] Tianxiang Bai and YF Li. Robust visual tracking with structured sparse representation appearance model. *Pattern Recognition*, 45(6):2390–2404, 2012.
- [11] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.
- [12] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

- [13] Emmanuel J Candes, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.
- [14] Harry Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3(1-2):111–133, 1956.
- [15] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- [16] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [17] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
- [18] Ingrid Daubechies, Jianfeng Lu, and Hau-Tieng Wu. Synchrosqueezed wavelet transforms: A tool for empirical mode decomposition. *arXiv preprint arXiv:0912.2437*, 2009.
- [19] Volker Roth and Bernd Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM, 2008.
- [20] David L Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.
- [21] Michael Elad, Mario AT Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.
- [22] Eric Lagasse, Heather Connors, Muhsen Al-Dhalimy, Michael Reitsma, Monika Dohse, Linda Osborne, Xin Wang, Milton Finegold, Irving L Weissman, and Markus Grompe. Purified hematopoietic stem cells can differentiate into hepatocytes in vivo. *Nature medicine*, 6(11):1229–1234, 2000.
- [23] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.
- [24] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.
- [25] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19:801, 2007.
- [26] Michal Aharon, Michael Elad, and Alfred Bruckstein. -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.
- [27] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm Computing Surveys (CSUR)*, 38(4):13, 2006.
- [28] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564–577, 2003.
- [29] Fatih Porikli, Oncel Tuzel, and Peter Meer. Covariance tracking using model update based on lie algebra. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 728–735. IEEE, 2006.
- [30] Shaohua Kevin Zhou, Rama Chellappa, and Baback Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *Image Processing, IEEE Transactions on*, 13(11):1491–1506, 2004.
- [31] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.
- [32] JingYu Yang, YiGang Peng, WenLi Xu, and QiongHai Dai. Ways to sparse representation: an overview. *Science in China series F: information sciences*, 52(4):695–703, 2009.
- [33] Baiyang Liu, Lin Yang, Junzhou Huang, Peter Meer, Leiguang Gong, and Casimir Kulikowski. Robust and fast collaborative tracking with two stage sparse optimization. In *Computer Vision—ECCV 2010*, pages 624–637. Springer, 2010.
- [34] Zhenjun Han, Jianbin Jiao, Baochang Zhang, Qixiang Ye, and Jianzhuang Liu. Visual object tracking via sample-based adaptive sparse representation (adasr). *Pattern Recognition*, 44(9):2170–2183, 2011.
- [35] Jie Shao, Nan Dong, and Minglei Tong. Multi-part sparse representation in random crowded scenes tracking. *Pattern Recognition Letters*, 2012.
- [36] Mert Dikmen and Thomas S Huang. Robust estimation of foreground in surveillance videos by sparse error estimation. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

Background Invariant Detection of Graphite in OMR Documents

Raghuveer Kanneganti¹, Randy Fry², Ahmed Fadhil¹, and Lalit Gupta¹

¹Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA

²McGraw-Hill Education, CTB, Monterey, CA, USA

Abstract - This study focuses on the development of a strategy to detect graphite responses in optical mark recognition documents using inexpensive visible light scanners. The main challenge in the formulation of the strategy is that the detection should be invariant to the numerous background colors and artwork in typical optical mark recognition documents. A test document is modeled as a superposition of a graphite response image and a background image. The background image in turn is modeled as superposition of screening artwork, lines, and machine text components. A sequence of image processing operations and a pattern recognition algorithm are developed to estimate the graphite response image from a test document by systematically removing the components of the background image. The image processing operations consist of gray-scale and color segmentation and the application of the Hough transform. Components that are not removed by image processing operations are identified and removed using a multivariate Gaussian classifier. The proposed strategy is tested on a wide range of scanned documents and it is shown that the estimated graphite response images are visually similar to those scanned by very expensive infra-red scanners currently employed for optical mark recognition. The robustness of the detection strategy is also demonstrated by testing a large number of simulated test documents.

Keywords: Optical mark recognition; texture classification; color segmentation

1 Introduction

The goal of this study is to introduce a strategy to detect graphite responses in scanned test documents using inexpensive scanners. The graphite responses of interest are the penciled bubbles and the alphanumeric characters written by the user in optical mark recognition (OMR) documents. OMR sheets are used extensively for evaluating the performances of students and for surveys. Currently, test documents are scanned with an infrared light source because most colored inks used for the background artwork are transparent to this wavelength while carbon absorbs the light. As a result, the background artwork vanishes leaving only the graphite responses and a few other carbon-bearing marks.

However, these OMR infrared scanners are expensive ($\approx \$100,000$) and are also quite expensive to maintain.

The detection of graphite responses through alternative methods is quite complex because there are hundreds of different types of optical mark recognition sheets which vary in background colors and artwork. Furthermore, optical mark recognition sheets may also be customized for specific needs. The graphite detection strategy described in this paper is aimed at significantly reducing the cost by using an inexpensive color scanner ($\approx \$75$) in conjunction with a sequence of image processing operations and a pattern classification algorithm. The goal is to obtain outputs that will be similar to those scanned by the very expensive OMR infrared scanners currently used.

To our knowledge, the specific problem addressed in this paper is quite unique. The development of OMR using an ordinary scanner has recently been reported in which tools are provided to design OMR sheets and to process the filled scanned sheets using knowledge of the design in the sheets [1]. The problem addressed in this paper, however, is quite different in the sense that the focus is on detecting graphite markings without any prior knowledge of the artwork design and colors in the OMR sheets. Other indirectly related studies which involve analyzing written pencil markings include hand writing recognition [2] and writer verification [3], however, they do not typically involve detection in backgrounds containing various artwork designs and colors.

2 Image Model

The output $F(x, y)$ of the inexpensive color scanner in response to a test optical mark recognition document is referred to as the “response image.” The response image can be modeled as the sum of two components: (1) the scanner response to the graphite, and (2) the scanner response to the blank optical mark recognition document which will be referred to as the background. If $G(x, y)$ and $N(x, y)$ are the graphite and background components, the response image is given by

$$F(x, y) = G(x, y) + N(x, y) \quad (1)$$

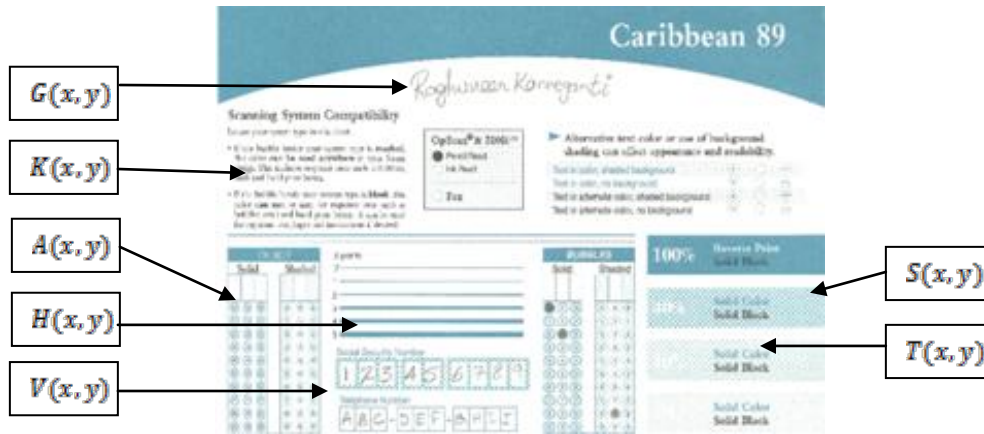


Fig. 1 Components of a typical OMR image

Using this model, the objective is to determine $G(x, y)$ given $F(x, y)$. If $N(x, y)$ is known and if $F(x, y)$ and $N(x, y)$ can be perfectly aligned, this problem could be easily solved by taking the difference between $F(x, y)$ and $N(x, y)$. In order to make the detection invariant to the numerous optical mark recognition backgrounds, it is assumed the blank optical mark recognition documents are not available, therefore, $N(x, y)$ is assumed unknown. Furthermore, perfect alignment in practice is impossible, therefore, it is not assumed. The approach developed in this paper systematically removes the background using a sequence of image processing and classification algorithms. The formulations of these algorithms are described in the following sections.

3 Image Process of a Response Image

Using this model, the objective is to determine $G(x, y)$ given $F(x, y)$. If $N(x, y)$ is known and if $F(x, y)$ and $N(x, y)$ can be perfectly aligned, this problem could be easily solved by taking the difference between $F(x, y)$ and $N(x, y)$. In order to make the detection invariant to the numerous optical mark recognition backgrounds, it is assumed the blank optical mark recognition documents are not available, therefore, $N(x, y)$ is assumed unknown. Furthermore, perfect alignment in practice is impossible, therefore, it is not assumed. The approach developed in this paper systematically removes the background using a sequence of image processing and classification algorithms. The formulations of these algorithms are described in the following sections.

$$N(x, y) = H(x, y) + V(x, y) + A(x, y) + S(x, y) + T(x, y) + K(x, y) \quad (2)$$

Figure 1 shows the components of a typical response image.

3.1 Horizontal and vertical line removal

The background components vary from document to document depending on the colors used to design the document. In order to make the detection of these components invariant to background color, the color response image $F(x, y)$ is converted into a gray scale image $f(x, y)$. The pixels of the solid vertical and horizontal lines in $f(x, y)$ can be detected by the colinearity Hough transform [4]. Moreover, because the pixels of the dot screening patterns also form horizontal and vertical lines, they may also be detected by the Hough transform and then removed from the background. In order to apply the Hough transform, the gray scale image is converted into a binary image $g(x, y)$ using a threshold that is determined autonomously as a function of the intensities of a given response. Let $A(\rho, \theta)$ be the colinearity Hough transform of $g(x, y)$ and let $\{A_{t_1}(\rho, 0)\}$ and $\{A_{t_2}(\rho, \pm 90)\}$ be the sets of accumulator cells with counts exceeding t_1 and t_2 , respectively. If $f_{t_1}(x, y)$ and $f_{t_2}(x, y)$ are the images formed by the pixels of $\{A_{t_1}(\rho, 0)\}$ and, $\{A_{t_2}(\rho, \pm 90)\}$, respectively, then the image given by

$$\tilde{g}(x, y) = g(x, y) - f_{t_1}(x, y) - f_{t_2}(x, y) \quad (3)$$

will be free of horizontal and vertical line segments with more than t_1 and t_2 pixels, respectively. The reason for imposing thresholds t_1 and t_2 is to make sure that short horizontal and vertical line segments which may be parts of graphite characters are not removed. The values of t_1 and t_2 can be determined empirically across a large number of response images.

3.2 Color-based segmentation

In the next step, $\tilde{g}(x, y)$ is used as a mask for removing the horizontal and vertical lines in the original color image. That is, the resulting color image is given by

$$\hat{F}(x, y) = F(x, y)\hat{g}(x, y) + A(x, y) + T(x, y) + K(x, y) + G(x, y) \quad (4)$$

As noted earlier, optical mark recognition sheets come in numerous colors. Therefore, the color clusters of the background images vary widely in the RGB color space. However, the cluster of the graphite component is relatively invariant in the RGB color space. To some degree (depending on the background colors), the removal of the background components can be made invariant to color by segmenting the graphite cluster in the RGB color space. The graphite pixels can be modeled as a (3×1) random vector Y whose density function is given by

$$p(Y) = \frac{1}{(2\pi)^{3/2}|\varepsilon|^{1/2}} \exp\{(Y - \mu)^T \varepsilon^{-1}(Y - \mu)\} \quad (5)$$

where μ and ε are the (3×1) and (3×3) mean and covariance matrices of Y , respectively. That is, the cluster is a hyper-ellipsoid. The major principal axis will be along the diagonal of the color cube which corresponds to the line connecting (0,0,0) to (255,255,255). The segmentation of the graphite cluster can, therefore, be expressed as

$$\hat{F}(x, y) = \begin{cases} 1, & \text{if } \hat{F}(x, y) < r^2 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where r^2 is the squared Mahalanobis distance [5] between Y and μ . The parameters μ and ε can be estimated from the graphite training set and the segmentation threshold r^2 can be determined as the smallest value that encompasses 100% of the training set cluster. The colored art components and the black text components of the background may or may not be partially removed in this step. In order for the formulation to be general, it is assumed that the segmented image is now

$$\hat{F}(x, y) = \hat{A}(x, y) + \hat{T}(x, y) + \hat{K}(x, y) + G(x, y) \quad (7)$$

where $\hat{A}(x, y)$, $\hat{T}(x, y)$, and $\hat{K}(x, y)$ are the modified components of $A(x, y)$, $T(x, y)$, and $K(x, y)$, respectively, due to color segmentation. The modified black machine text component $\hat{K}(x, y)$ is relatively dark and can be removed using intensity based segmentation. The resulting gray scale image $h(x, y)$ may now be expressed as

$$h(x, y) = h_A(x, y) + h_T(x, y) + h_G(x, y) \quad (8)$$

where, $h_A(x, y)$, $h_T(x, y)$, and $h_G(x, y)$ are the gray scale components corresponding to $A(x, y)$, $T(x, y)$, and $G(x, y)$, respectively.

4 Classification of Components

Due to the overlap of the clusters in $h(x, y)$, the removal of $h_A(x, y)$ and $h_T(x, y)$ directly using image processing

algorithms is not possible, therefore, the components of $h(x, y)$ are separated using a texture recognition approach which is used in numerous problems related medical image analysis [6], remote sensing [7], and automated inspection [8]. Texture is chosen here because the graphite component is expected to have a coarser texture when compared with the other two machine printed components. Further more, texture carries over from the color domain to the gray-scale domain thus simplifying the detection of texture in one gray-scale image rather than in three (red, green and blue) images. Therefore, $h(x, y)$ is considered to be an image containing objects belonging to two classes ω_1 (graphite) and ω_2 (non-graphite). The texture features are selected according to the following criteria: they must be invariant to the object (a) shape because the objects within each class have varying shapes, and (b) size because the objects within a class have varying dimensions. In order to satisfy these requirements, texture features are derived from the normalized histograms of the objects because the normalized histogram, represented by $p(z_i)$, $i = 1, 2, \dots, 255$, is invariant to the shape and size of the objects. The following features are selected:

Entropy:

$$e(z) = -\sum_{i=0}^{255} p(z_i) \log_2 p(z_i) \quad (9)$$

Skewness:

$$\mu_2(z) = \sum_{i=0}^{255} (z_i - m)^2 p(z_i) \quad (10)$$

Uniformity:

$$U(z) = \sum_{i=0}^{255} p^2(z_i) \quad (11)$$

Using the above feature set, a 3-dimensional multivariate classifier whose discriminant function is given by [9]

$$d_i(Z) = \ln P(\omega_i) - \frac{1}{2} \ln |\varepsilon_i| - \frac{1}{2} [(Z - \mu_i)^T \varepsilon_i^{-1} (Z - \mu_i)], \quad i = 1, 2 \quad (12)$$

is designed to separate the two classes. The parameters μ_i and ε_i are the (3×1) mean vector and (3×3) covariance matrices of the classes labeled ω_1 and ω_2 , respectively. A test object represented by a vector Z^* is assigned to class ω^* given by

$$\omega^* = \arg \max_i [d_i(Z^*)]. \quad (13)$$

5 Detection Experiments and Results

Experiments were designed to subjectively and objectively evaluate the performance of the graphite detection procedure developed in this paper. The objective evaluations were conducted by visually comparing the detected output images with the corresponding OMR output images of the same test sheets. The probability of error between detected outputs of simulated response images and a known graphite image was used as an objective measure of the performance.

5.1 Subjective evaluations

The data set used to design and evaluate the procedure for this case consisted of a wide range of optical mark recognition sheets with different background colors and artwork. Each sheet had penciled bubbles and penciled characters written by different individuals using different graphite pencils. The spatial resolution of the scanner was set to 300 dots per inch. In order to evaluate the performance of the proposed strategy, the detected outputs of the proposed strategy were compared, visually, with OMR outputs of the same sheet using an OpScan iNSIGHT 4 infra-red scanner. The graphite training set was generated by selecting penciled pixels from a scanned image of a white sheet with penciled characters (see Figure 2). Pixels of the scanned "Stormy" optical mark recognition sheet, shown in Figure 3, were selected to form the training set of the non-graphite class. For brevity, only two examples of the scanned test sheets used in the experiments and the corresponding OMR scanned outputs and the detected outputs are shown in Figure 4. The results clearly show that the detected output images and OMR output images are quite similar. Moreover, these results are typical of the results obtained for the other test sheets.

5.2 Objective evaluations

For this case, the data set consisted of simulated response images generated by the following model:

$$F(x, y) = G(x, y) + N(x, y) + \eta(x, y) \quad (14)$$

where $G(x, y)$ is the graphite image, $N(x, y)$ is the background (blank optical mark recognition) image, and $\eta(x, y)$ is the noise introduced during scanning. The scanning noise component is included to account for the typical variations that occur when a given image is scanned repeatedly. That is, each pixel at a given location varies randomly from scan-to-scan. The scanning noise is assumed to be zero-mean Gaussian with variance σ^2 . A specific response $F_i(x, y)$ is given by

$$F_i(x, y) = G(x, y) + \alpha_i F_{R_i}(x, y) + \beta_i F_{G_i}(x, y) + \gamma_i F_{B_i}(x, y) + \eta(x, y) \quad (15)$$

where, α_i , $0 \leq \alpha_i \leq \left(\frac{255}{m_r}\right)$; β_i , $0 \leq \beta_i \leq \left(\frac{255}{m_g}\right)$; and γ_i , $0 \leq \gamma_i \leq \left(\frac{255}{m_b}\right)$, are the coefficients that determine the color of the resulting background image. The maximum values of the red, green, and blue component images of $N(x, y)$ are represented by m_r , m_g , and m_b , respectively.

For the experiments, the simulated responses were generated using the following sequence of steps: (i) One blank optical mark recognition sheet was selected and scanned. This gave the image $N(x, y)$; (ii) A small smooth region of constant intensity in $N(x, y)$ was chosen and the variance of the pixels in this region was estimated. This estimate was used for the scanning noise variance σ^2 ; (iii) Ten individuals with different

pencils wrote and filled bubbles in various parts of the selected optical mark recognition sheet and the resulting penciled sheet was scanned;

(iv) A simulated response image was obtained by changing the color coefficients of $N(x, y)$ randomly, adding Gaussian noise with zero mean and variance σ^2 , and adding the graphite image $G(x, y)$ to the result.

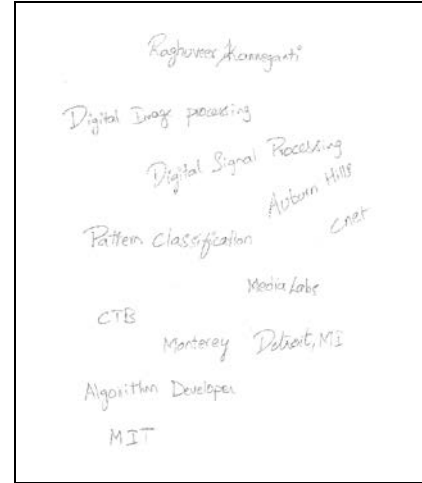


Fig.2 Image used to extract graphite in the training set

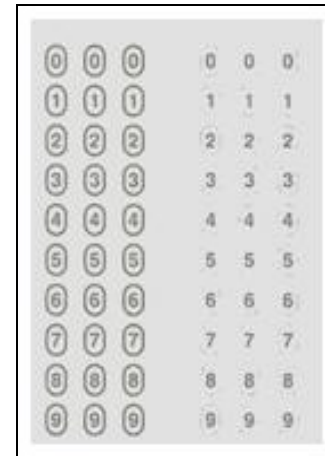


Fig.3 Image used to extract non-graphite in the training set

It should be noted that numerous response images with a multitude of background colors can be generated using the method outlined above. Another advantage is the fact that a single image $G(x, y)$ encompasses the writings of several individuals using different pencils. Most importantly, the performance of the graphite detection methodology can be evaluated objectively over a large number of response images. The two types of detection errors that can occur are (a) (b/g) : a graphite pixel misclassified as a background pixel, and (b) (g/b) : a background pixel misclassified as a graphite pixel. Therefore, for a given response image, the detection error probability can be determined from

$$P(E) = P(b/g)P(g) + P(g/b)P(b) \quad (16)$$

The prior probabilities $P(g)$ and $P(b)$ of the graphite and background pixels can be estimated from a response image and the detection results can be used to estimate the error probabilities.

For the detection experiments, the graphite training set, the segmentation factor, and the Hough transform thresholds were the same as those used in the subjective evaluation experiments. Figure 5 shows two examples of simulated response images and the detection results for these images. The corresponding detection errors are listed in the figure captions. Error probabilities of such magnitudes show very little visual difference between the detected outputs and the

graphite image. Figure 6 shows the detection error probabilities of 125 simulated response images. Each notch on the horizontal axis represents a test response image and the vertical axis shows the detection error probabilities. The average error across all 125 response images was 0.00304. From the subjective and objective results, it can be concluded that the proposed strategy is quite effective in estimating the graphite response in test documents. It is also important to note that documents with widely varying colors were tested; however, the training sets for each class were derived from single-images.

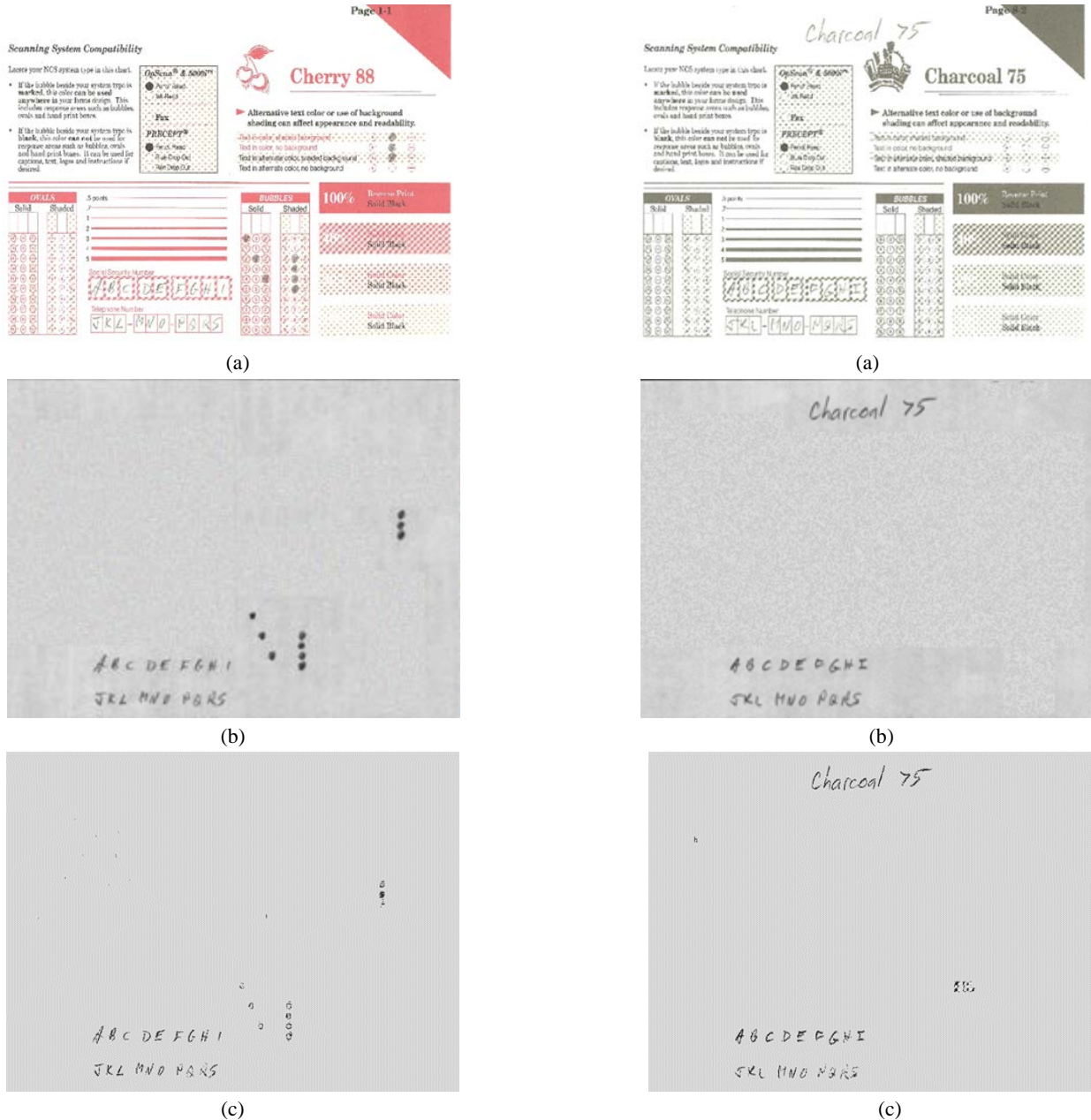


Fig. 4 Examples of (a) response image (b) OMR scanned output (c) detected output



(a)



(b)

Fig. 5 Examples of simulated response images and the detection results for these images. The corresponding detection error probabilities are (a) 0.0032 (b) 0.0028

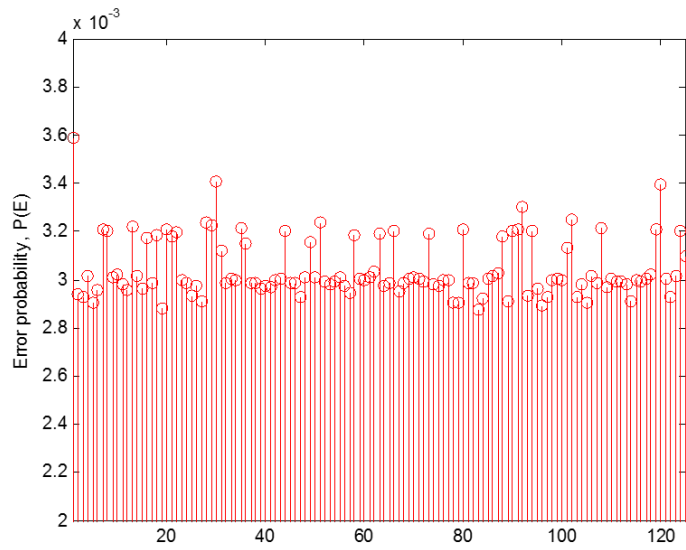


Fig. 6 Detection error probabilities of 125 simulated response images

6 Conclusions

This study focused on introducing a strategy to detect the graphite components in scanned test documents with varying background colors and art work. A test image was modeled as a superposition of a graphite image and the background image. The background image was further modeled as a superposition of several components. The estimation problem was formulated in terms of background component removal and a sequence of operations was developed to systematically remove the components. The strategy was applied to a wide range of test documents scanned using an inexpensive scanner and it was shown that results were visually similar to those obtained using an expensive OMR scanner. The robustness of the detection strategy was also demonstrated by testing a large number of simulated test documents.

7 Acknowledgements

This research project was supported by CTB/McGraw Hill LLC under a grant entitled "Color-Agnostic Dropout of Document Backgrounds." The authors would like to express their sincere thanks to Michelle Boyer for her invaluable support during the entire project. The opinions and conclusions contained in this paper are solely those of the author and do not necessarily reflect the policy or position of CTB/McGraw-Hill. The methods described in this paper are patent pending.

8 References

- [1] Garima Krishna, Hemant Ram Rana, Ishu Madan, Kashif, Narendra Sahu, 'Implementation of OMR technology with the help of ordinary scanner,' International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 4, April 2013
- [2] Rejean Plamondon, Sargur N. Srihari, 'On-Line and Off-Line handwriting recognition: A comprehensive survey,' IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 22, Issue 1, January 2000
- [3] R. K. Hanusiak, L.S. Oliveria, E. Justino, R.Sabourin, 'Writer based verification using texture-based features,' IJDAR 2012
- [4] Duda, R. O. and P. E. Hart, 'Use of the Hough Transformation to Detect Lines and Curves in Pictures,' Comm. ACM, Vol. 15, pp. 11-15, January, 1972
- [5] Dougherty G, 'Pattern Recognition and Classification: An Introduction,' Springer, New York, 2013
- [6] Eun-Byeol J, Ju-Hwan L, Jun-Young P, Sung-Min K, 'Detection of breast cancer based on texture analysis from digital mammograms,' Intelligent Autonomous Systems12, AISC 194, pp. 893-900, 2013
- [7] Mahmoud A, Elbially S, Pradhan B, Buchroithner M, 'Field-based land cover classification using TerraSAR-X texture analysis,' Advances in space research, 48(5), pp. 799-805, 2011
- [8] Xie X, 'A review of recent advances in surface defect detection using texture analysis techniques,' Electronic letters on computer vision and image analysis, Vol. 7, Issue 3, pp. 1-22, 2008
- [9] Duda R, Hart P, Stork D, 'Pattern Classification,' second edition, A Wiley-Interscience publication, New York, 2001

Investigation of Using the Local Directional Pattern for stereo corresponding problem

Fadl Dahan, Khalid Al-Mutib, Mansour Alsulaiman, Mohammed Faisal, and Ramdane Hedjar
College of Computer and Information Sciences
King Saud University
Al-Riyadh, Saudi Arabia
{fadhl, muteb, msuliman, mfaisal, hedjar, }@ksu.edu.sa

Abstract

Despite the significant success of using local feature descriptor with many applications such as, face detection, human facial expression recognition, gender classification, signature verification, etc. Based on our knowledge, there is no research present on the usage of the descriptor to solve the stereo corresponding problem. Despite this, descriptor uses the same methodology for comparing the images blocks. In this paper, we developed a new method to solve the stereo vision corresponding problems by using the Local Directional Pattern (LDP). The proposed method encodes the input images using LDP encoder, then for generating depth map, we use block-matching algorithm but we change distance measurement (chi-square) instead of similarity measurements to compare the blocks code histogram. The proposed method shows promising results comparing with other methods on testing images benchmark.

Key words: stereo corresponding, depth map, block matching, local feature descriptor, local directional pattern (LDP).

1. Introduction

The backbone of stereo vision applications based is the accurate disparity image so there are many efforts to get an accurate estimation for disparity image. In stereo vision corresponding problem, we aim to obtain the depth information from two or more digital images of the same scene. Simple stereo system utilizes two images, these images are taken from two parallel cameras placed horizontally with distance called "baseline". The output of stereo analysis is depth map shows how far the points from the camera in the physical scene. However, this field suffers from the accuracy and computation time problems[1], so many algorithms proposed for fixing these problems.

We can categorize the proposed algorithm into two main types Area Based Algorithm (ABA) and Feature Based Algorithm (FBA). ABA creates the disparity map based on the correlation using the pixels of the local neighborhood. Additionally, the pixel intensity of the blocks are used to compare between the two stereo images. We start this process by determine the interest pixels and take their neighborhood, then we compare this area with same area in the compared image based on overlapping pixels (disparity range). The problems behind this method are the time consuming, and the sensitivity of the noise and distortion, because it based on the pixels intensity [2].

FBA solves the matching process by searching for intensity feature from the image to compare it with other image. Mostly, the FBA used for feature edge point or edge segment [3]. The advantages of this type are the simplicity comparison process, and the execution time (it's faster than area based algorithms).

Efficient methods such as local image descriptor is used in many applications such as, face recognition, texture classification, human computer interaction, gender classification. According to these applications, we release that, the local image descriptor can be used in the disparity image construction. Methods under local image descriptor search for a way to describe the images in different condition like lighting, rotation. This searching process divided into two groups [4]. The first group called sparse descriptor. The sparse descriptor searches for interest point in the image, and takes the sampling image patch around this point to use it in matching process. SIFT (Scale Invariant Feature Transform) [5] and its variants are an example of sparse descriptor. Second group called dense descriptor. The dense descriptor searches pixel by pixel over the given image to extract the local features [6]. Local binary pattern (LBP) [7], Gabor [8], and LDP [6] are an examples of dense descriptor.

The idea behind LDP descriptor is that, it takes each pixel from the image with eight directions and computes the response between each pixel and the directions. After that, the result encoded from the relative strength magnitude. LDP successfully applied in many applications, such as object recognition [6], face recognition [9], gender classification [10], facial expression recognition [11, 12], and signature verification [13].

We organized this paper as following: related work is presented in section 2. In section 3, we explained the proposed method. The experimental results are presented in section 4. Finally, the conclusion is presented.

2. Related work

Based to our knowledge, our research is the first research, which uses the descriptor to solve the stereo corresponding problem. In this section, we are going to present some of the bases and famous corresponding algorithm. Basic Block Matching BBM is an approach that used to find the corresponding in the images in stereo vision. The main idea of BBM is to find the similarity by comparing the intensity of two blocks from input images. Many methods are classified as a BBM such as, Sum of Absolute Differences (SAD) [14, 15], Sum of Square Difference [14], Zero-mean Sum of Absolute Differences (ZSAD) [16], and Normalized Cross Correlation (NCC) [16]. However, BBM suffers from the noisy and accuracy [1]. Therefore, we can improve it by applying another vision approaches such as Sub-pixels Estimation SE, Dynamic Programing DP, and Image Pyramiding (IP). SE uses to smooth the different disparity between regions, where in the block matching we only use the minimum difference between the blocks. However, with sub-pixels the minimum difference comes from the minimum difference and two neighboring minimum distance [17]. The problem of depth map comes from finding the optimal disparity between right and left images where in block

matching the optimal disparity comes from the cost function for each pixel alone. DP introduced another way to find the optimal disparity, where optimal path (actual disparity) is calculated between right and left images corresponding pixels by using the location difference [2, 18]. The main disadvantage of the BBM and DP is the time consuming. IP introduce solution for corresponding problem by reducing the size of the image in each level by half [1].

Recently, research in local texture encoding utilizes the gradient magnitude change between the neighbored pixels in all directions instead of comparing the intensity of neighboring pixels like LBP. Consequently, the local texture encoding is able to analysis different magnitude in all directions to encode particular pixel. According to this, authors of [6] proposed their novel descriptor called Local Directional Pattern (LDP). LDP encodes input image by assigning an eight binary for each pixel bit. This process is computed by comparing the central pixel value in different direction [6, 9]. For that, authors used Kirsch masks to compute the edge response value for each pixel among eight different orientations (M_0 - M_7). Figure 1 illustrates these masks.

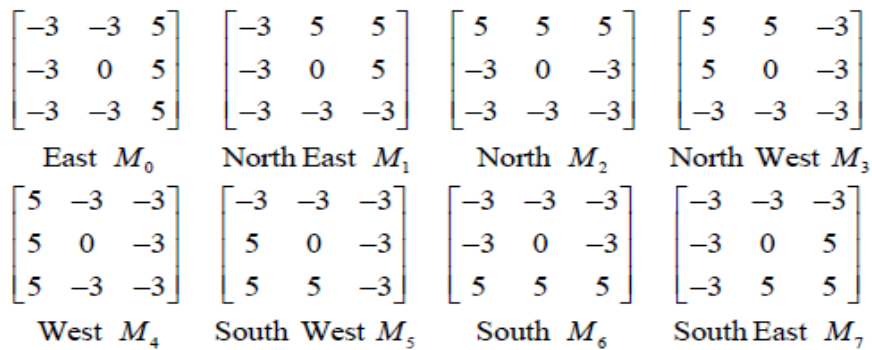


Figure 1: Eight directions Kirsch edge masks

The encoding process of the input image start by dividing image into 3×3 blocks, then for each block, we use the masks to obtain the edge response m_0, m_1, \dots, m_7 [6]. The obtained values are the eight directions for centric pixel. The encoding of the LDP is based on the prominent direction k , which its values become 1 and others values $(8-k)$ become 0. Figure 2 shows an example of the LDP encoding process.

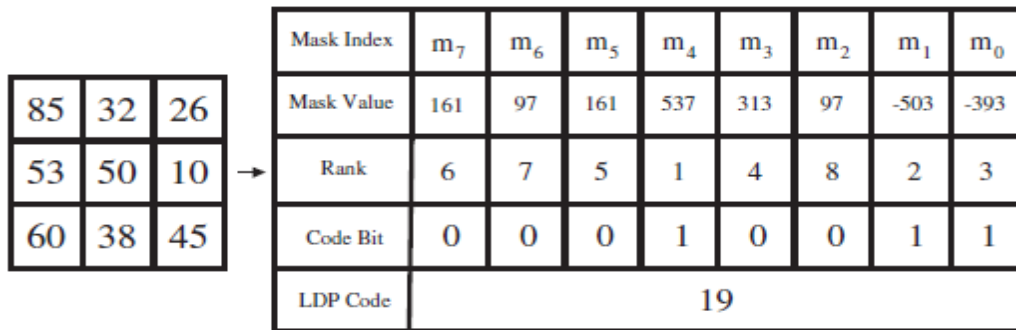


Figure 2: LDP encoding example [10]

3. Proposed method

The proposed method, we construct the depth image from left and right images by using the disparity image algorithm and local image descriptor. Our contribution comes from encoding the input images using the Local Directional Pattern (LDP), then generating the histogram code instead of the pixel intensity. After that, using the chi-square distance measurement instead of similarity with block matching to find the stereo corresponding. Figure 3 shows the process model (LDP Encoding, Histogram Generation, and Stereo Corresponding) of the proposed method.

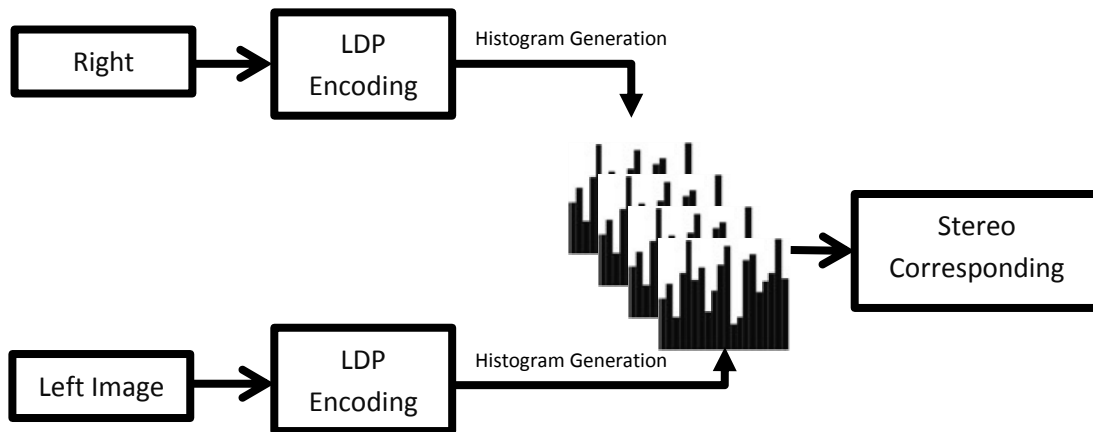


Figure 3: Proposed method

a) LDP Encoding Process

The LDP encoding process responsible for encoding the input images, to be able to use the distance measurement. The LDP encoding process encodes the input images and generates the encoded image pixels, which will use in next process. We assume that, the input images are in gray scale and already rectified. The LDP encoder takes each input image and divides it into 3×3 blocks.

For each block, we multiply it by all masks, and then rank the results. The three highest response values ($k=3$) become 1 and the other 0. Equation 1 shows the encoding process. Figure 4 illustrates this process.

$$LDP_k = \sum_{i=0}^7 b_i(m_i - m_k) \times 2^i \quad (1)$$

$$b_i(a) = \begin{cases} 1 & a \geq 0 \\ 0 & a < 0 \end{cases}$$

where m^k is the k^{th} most significant directions and $k=3$

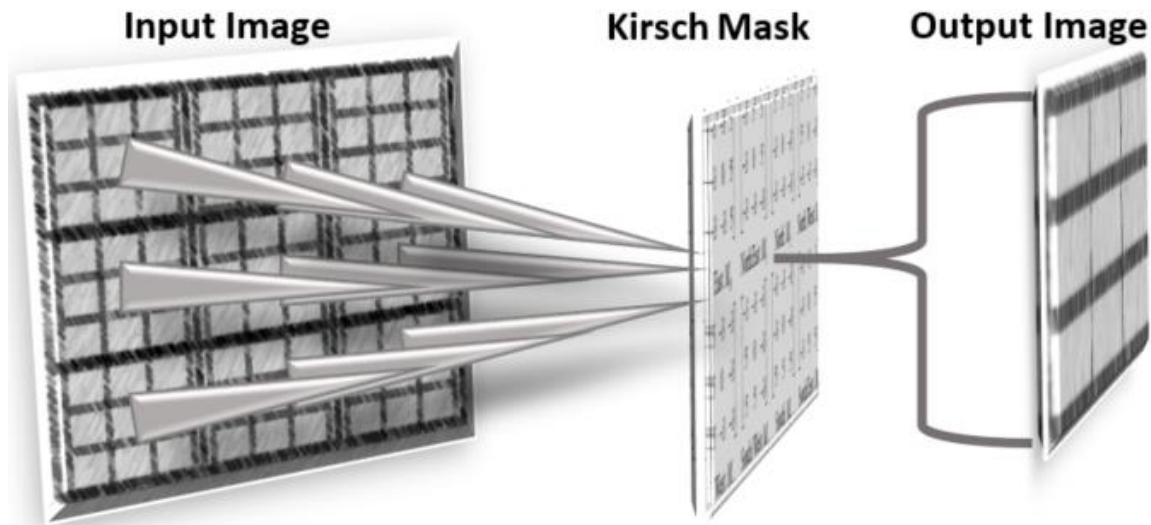


Figure 4: LDP encoding process

b) Histogram Generation Process

The output of the first process (the encoded image pixels from LDP), give us the ability to use the distance measurement with block matching to get the depth image. Here, we use the distance measurement as in face recognition instead of using similarity measurements. In this step, we take a block from left and right images and generate its histogram; these blocks have the same size and same number of bins. In the proposed method, the size of left and right blocks must be larger than similarity measurement. This due to accuracy of the histograms, the accuracy becomes better with a large block, and leads to better depth image generation. The output of this process is the histogram of the left and right blocks.

c) Stereo Corresponding Process

There are two measures of the cluster analysis between two objects, the indicating similarity and indicating distance [19]. In stereo corresponding process, we take the left and right histograms and apply the Chi-square equation to find the distance between the histograms. Based on the disparity range, we calculate the distance between the one histogram from the right image and n histograms from the left image, in contrast with these similarity measurements. In our method, we use distance measurements (Chi-square) between the blocks pixel histogram, which are generated from encoding the images using LDP descriptor in the histogram generation step.

$$\chi^2(x, y) = \frac{\sum_{i=1}^n \left(\frac{(x_i - y_i)^2}{x_i + y_i} \right)}{2} \quad (2)$$

4. Experimental

In order to test the proposed method, we used four images from Middlebury benchmark database. These images are Tsukuba, Venus (simple geometry), Teddy, and Cones (complex geometry). Based on our knowledge, there is no quantitative estimation to evaluate the disparity image, which is used in the proposed method, because it is very difficult and the evaluation is done by using subjective visual estimation [2]. Even though, in this work, we evaluate our disparity images by quantitative estimation proposed by [2]. However, this evaluation is not a direct evaluation.

Figure 5 shows the output depth images for one stereo image (Tsukuba) generated by the proposed method, and compared with the other three methods mentioned in the literature. We have taken care about tuning the parameters such as block size and disparity range to give best results for each algorithm because the fixed setting doesn't work for all.

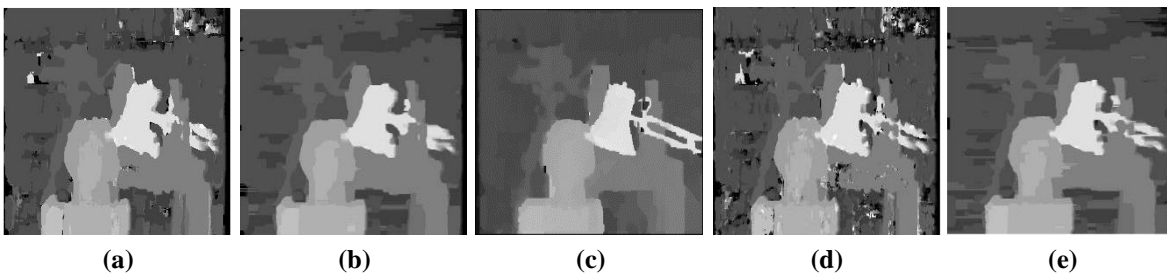


Figure 5: The depth images for Tsukuba image generated using: (a) Block matching with SAD, (b) Dynamic programming, (c) Proposed method, (d) Sub-pixel estimation and (e) Pyramiding.

In this quantitative evaluation, we compare the reconstructed disparity images with the ground-truth disparity image (attached with the images benchmark). The peak signal-to-noise ratio measuring used:

$$PSNR = 10 \log_{10} \left(\frac{255 * 255}{MSE} \right) dB,$$

Where

$$MSE = \frac{1}{k_1 k_2} \sum_{i=0}^{k_1-1} \sum_{j=0}^{k_2-1} \left(f_i(i, j) - \tilde{f}_i(i, j) \right)^2$$

Where

k_1 : The image size of reconstructed disparity image.

k_2 : The image size of idle disparity image.

$f_l(i, j)$: The reconstructed disparity image.

$\tilde{f}_l(i, j)$: The ground truth for input image.

To test the proposed method, we compare the results with four algorithms shown in table 1 and use the Middlebury benchmark database.

Images	Basic-Block Matching	Dynamic Programing	Proposed Method	Sub-pixel Estimation	Pyramiding
Tsukuba	58.5680	58.3084	58.1718	57.7980	58.0150
Venus	62.6736	61.8044	61.8515	61.7283	60.8301
Teddy	56.5550	56.6178	57.9828	57.1297	58.2095
Cones	57.9992	58.1028	58.1662	57.2625	59.3135

Table 1: comparing our result with other methods

As we can see in table 1, our method performance is comparable (Slightly better) with basic-block matching and dynamic programing on simple geometry images (Tsukuba and Venus), while pyramiding and sub-pixel estimation do better. For complex geometry images, the results are varied and differentiated, but the proposed method performance is still good.

5. Conclusion

In this paper, we proposed a simple and robust corresponding method inspired by local feature descriptor. We investigate applying LDP descriptor to get an accurate estimation for disparity image. In this research, we replace similarity measurements by distance measurements with block matching to get an accurate estimation. The proposed method shows encouraging results, and let us says that. We can use face recognition, gender classification, and other methods with stereo corresponding problem. Solving corresponding problem by using these methods is promising, but needs more research. Experimental results show that our method is comparable with famous and basis corresponding algorithms especially with block matching with similarity measurements.

References

- [1] A. Koschan, V. Rodehorst, and K. Spiller, "Color stereo vision using hierarchical block matching and active color illumination," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, 1996, pp. 835-839.
- [2] C. S. Park and H. W. Park, "A robust stereo disparity estimation using adaptive window search and dynamic programming search," *Pattern Recognition*, vol. 34, pp. 2573-2576, 2001.
- [3] U. R. Dhond and J. K. Aggarwal, "Structure from stereo-a review," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, pp. 1489-1510, 1989.

- [4] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 1705-1720, 2010.
- [5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91-110, 2004.
- [6] T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (LDP)—A robust image descriptor for object recognition," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 2010, pp. 482-487.
- [7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, pp. 2037-2041, 2006.
- [8] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, pp. 837-842, 1996.
- [9] T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (LDP) for face recognition," in *Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference on*, 2010, pp. 329-330.
- [10] T. Jabid, M. Hasanul Kabir, and O. Chae, "Gender classification using local directional pattern (LDP)," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 2162-2165.
- [11] T. Jabid, M. H. Kabir, and O. Chae, "Facial expression recognition using local directional pattern (LDP)," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010, pp. 1605-1608.
- [12] T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI journal*, vol. 32, pp. 784-794, 2010.
- [13] M. A. Ferrer, F. Vargas, C. M. Travieso, and J. B. Alonso, "Signature verification using local directional pattern (LDP)," in *Security Technology (ICCST), 2010 IEEE International Carnahan Conference on*, 2010, pp. 336-340.
- [14] M. J. Hannah, "Computer matching of areas in stereo images," DTIC Document 1974.
- [15] S. T. Barnard, *A stochastic approach to stereo vision*: Defense Technical Information Center, 1986.
- [16] J. Banks, M. Bennamoun, and P. Corke, "Non-parametric techniques for fast and robust stereo matching," in *TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE*, 1997, pp. 365-368.
- [17] E. Trucco and A. Verri, *Introductory techniques for 3-D computer vision* vol. 93: Prentice Hall Englewood Cliffs, 1998.
- [18] O. Veksler, "Stereo correspondence by dynamic programming on a tree," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 384-390.
- [19] C. R. Rao and T. S. Rao, *Handbook of statistics*: Elsevier, 2006.

Preliminary study on the hand clapping action recognition based on the Leap Motion Controller

Jin-Woo Jung¹, Hyeseong Lee², Jung-Soo Park³ and Yong-One Cho⁴

^{1,2,3,4}Department of Computer Science and Engineering, Dongguk University, Seoul, Korea

¹Email : jwjung@dongguk.edu (corresponding author)

Abstract - In this paper, we propose a method of recognizing human hand clapping action based on the Leap Motion Controller. Basic hand clapping is a kind of simple but somewhat fast periodic hand action. By the inevitable occlusion of two hands, fast speed of hand motion, and its neighboring noisy data, it is not easy to be recognized either by vision sensors or by sound sensors.

Leap Motion Controller (LMC) is one of the most advanced and low-priced devices which can be used to detect the detailed motion of multiple moving objects in some restricted range. By the fast and accurate hand trajectory data from Leap Motion Controller, we can predict the relative position of two palms more accurately. And, by using the prior knowledge about basic hand clapping action sequence, LMC-based hand clapping recognizer has been developed. The average success ratio is 86.11% with the three different test users and 240 tests per each user.

Keywords: hand clapping, action recognition, Leap Motion Controller

1 Introduction

Nowadays, gesture recognition is one of the main issues in the human-computer interaction researches and developments [1]. As results, commercial low-priced sensing devices such as Microsoft Kinect have been developed and widely used. But, it is not easy to make some recognizer for fast and accurate motion since Microsoft Kinect is based on the image processing technique and undergoes some problems such as occlusion and noisy data [2]. Leap Motion Controller (LMC) is one of the most advanced and low-priced devices which can be used to detect the detailed motion of multiple moving objects in some restricted range [3,4]. By the fast and accurate hand trajectory data from LMC, we can predict the relative position of two palms more accurately.

Basic hand clapping is a kind of simple but somewhat fast periodic hand action. In addition, by the inevitable occlusion of two hands, fast speed of hand motion, and its neighboring noisy data, it is not easy to be recognized either by vision sensors or by sound sensors.

In this paper, a simple algorithm is addressed to make a hand clapping action recognizer based on LMC. Some information on Leap Motion Controller is summarized in the section 2. And in the section 3, overall structure of LMC-based hand clapping recognizer is introduced. Finally, its experimental results are shown in the section 4.

2 Leap Motion Controller

Leap motion controller (LMC) is a sensor device that was developed by Leap Motion, Inc. It can recognize 3D (x,y,z) locations, gestures, and motions of various objects including human hand and fingers. The measurement is made by the coordinates such as Fig.1 and the unit of output value is mm. The recognition range of LMC is from 25 to 600 mm, shaping as an inverted pyramid [4].

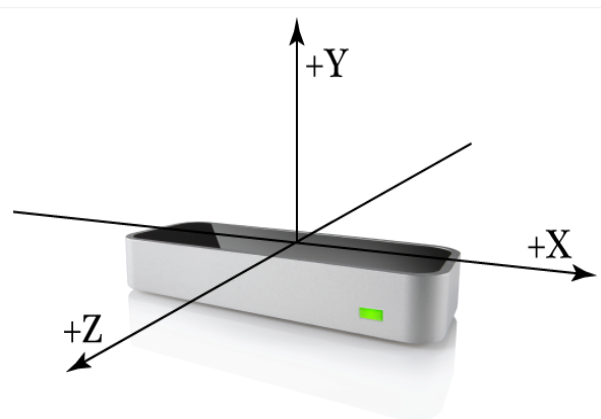


Figure 1. Coordinates of Leap Motion Controller [4]

The recognition of hand or finger motion is started from the camera frame capturing and the motion information of hand and fingers are represented as its own tracking data as follows [4].

1) Frame list and Frame motion

(1) Frame List of tracking data

- Hands - All hands
- Pointables - All fingers and tools as Pointable Objects
- Fingers - All the fingers
- Tools - All the tools
- Gestures - All the gestures that started, ended, or which had as update

(2) Frame motion

The frame motion information provided by Leap API are as follows:

- Rotation Axis - A direction vector expressing the axis of rotation
- Rotation Angle - The angle of rotation clockwise around the rotation axis (using the right-hand rule)
- Rotation Matrix - A transform matrix expressing the rotation
- Scale Factor - A factor expressing expansion or contraction
- Translation - A vector expressing the linear movement

2) Hand model and Hand attributes

The hand attributes provided by Leap API are as follows:

- Palm Position - The center of the palm measured in millimeters from the Leap Motion origin
- Palm Velocity - The speed of the palm in millimeters per second
- Palm Normal - A vector perpendicular to the plane formed by the palm of the hand. The vector points downward out of the palm
- Direction - A vector pointing from the center of the palm toward the fingers
- Sphere Center - The center of a sphere fit to the curvature of the hand (as if it were holding a ball)
- Sphere Radius - The radius of a sphere fit to the curvature of the hand. The radius changes with the shape of the hand

3) Hand motion

The hand information provided by Leap API are as follows:

- Rotation Axis - A direction vector expressing the axis of rotation
- Rotation Angle - The angle of rotation clockwise

- around the rotation axis (using the right-hand rule)
- Rotation Matrix - A transform matrix expressing the rotation
- Scale Factor - A factor expressing expansion or contraction
- Translation - A vector expressing the linear movement

4) Finger and Tool lists

The finger and tool information provided by Leap API are as follows:

- Length - The length of the visible portion of the object (from where it extends out of the hand to the tip)
- Width - The average width of the visible portion of the object
- Direction - A unit direction vector pointing in the same direction as the object (i.e. from base to tip)
- Tip Position - The position of the tip in millimeters from the Leap Motion origin
- Tip Velocity - The speed of the tip in millimeters per second

5) Gestures

Leap API provides the following 4 gestures as APIs.

- A circle gesture with the forefinger
- A horizontal swipe gesture
- A key tap gesture with the forefinger
- A screen tap gesture with the forefinger

Whenever LMC detects some object, Leap Motion Software gives it a unique identification number. And this identification number is maintained during the time period just when the object is visible.

3 LMC-based hand clapping recognizer

Human hand clapping can be understood as an action of fast, periodic and symmetric property like Fig. 2.

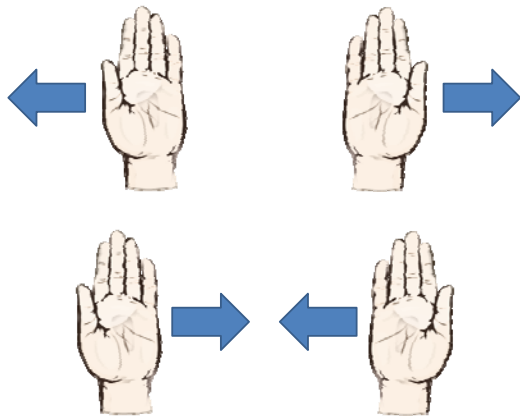


Figure 2. Human hand clapping

To design LMC-based hand clapping recognizer more simply, two assumptions are used as follows:

[Assumptions]

1. There are always two hands (left hand and right hand) in the sensing range of LMC.
2. There are always more than two clapping events in human hand clapping action. As a result, there is always some time when two hands are going in the opposite directions during human clapping action.
3. During human clapping action, the trajectory of hand movement is located on the same line.
4. During human clapping action, there is always some time when the distance between two hands is near zero.

Based on the above assumptions, the algorithm for LMC-based hand clapping recognizer are as follows:

[Algorithm]

Input: sequence of human hand actions

Output: One of the following events,
 {Clap-Hands are clapped,
 Moving Away-Hands are going away each other,
 Moving Closer-Hands are going closer each other}

Procedure:

1. Find the positions of left palm and right palm from LMC palmPosition() function
2. Calculate the motion vector of left palm and right palm using the current palm positions and the previous palm positions

3. Calculate the angle between left palm's motion vector and right palm's motion vector
4. If the difference between the angle from step 3 and 180 degree, Ang_{palm} , is bigger than the threshold angle value, Th_{angle} , go to step 1
5. Calculate the current palm distance between left palm and right palm
6. Wait until the current palm distance from step 5, $Dist_{palm}$, is bigger than the previous palm distance. If there is such an event that $Dist_{palm}$ is bigger than the previous palm distance, output 'Moving Away-Hands are going away each other' and go to step 7. If there is no such event until the threshold wait time, Th_{wait} , then go to step 1. During the waiting time, output 'Moving Closer (Hands are going closer each other)' whenever the current palm distance, $Dist_{palm}$, is smaller than the previous palm distance.

Wait until the current palm distance, $Dist_{palm}$, is less than the threshold distance value, Th_{dist} . If there is such an event until the threshold wait time, output 'Clap (Hands are clapped)' and go to step 1. If there is no such event until the threshold wait time, Th_{wait} , then go to step 1. During the waiting time, output 'Moving Closer (Hands are going closer each other)' whenever the current palm distance, $Dist_{palm}$, is smaller than the previous palm distance, and output 'Moving Away-Hands are going away each other' whenever the current palm distance, $Dist_{palm}$, is bigger than the previous palm distance.

4 Experimental results

4.1 Experimental Settings

To verify the effectiveness of the proposed hand clapping recognition method, three different test subjects participated in these recognition experiments and 240 tests are performed per each test subject.

Fig.3 shows the sensing range of LMC. Based on the X-axis of LMC sensing range, we partitioned the sensing range into six different sub-regions designating for the target position of hands collision. In Fig.4, sub-region ④, ⑤, and ⑥ are located 25cm high from the floor and sub-region ①, ②, and ③ are located 40cm high from the floor. ②, ⑤ are located in the center region. ①, ④ are located 20cm left from the center point, i.e., -20cm in X-axis. ③, ⑥ are located 20cm right from the center point, i.e., +20cm in X-axis.

Each test subject did the hand clapping recognition experiments 20 times per each sub-region for hands collision. In each test trial, each test subject is required to do his hands clapping targeted at the center point of each sub-region. Additionally, two different threshold value sets are used for

each experiment to check the effect of threshold values. The first threshold value set was determined as $Th_{angle} = 180 \times 0.2$ (deg), $Th_{dist} = 40$ (pt), $Th_{wait} = 1$ (sec). And the second threshold value set was determined as $Th_{angle} = 180 \times 0.3$ (deg), $Th_{dist} = 50$ (pt), $Th_{wait} = 1$ (sec).

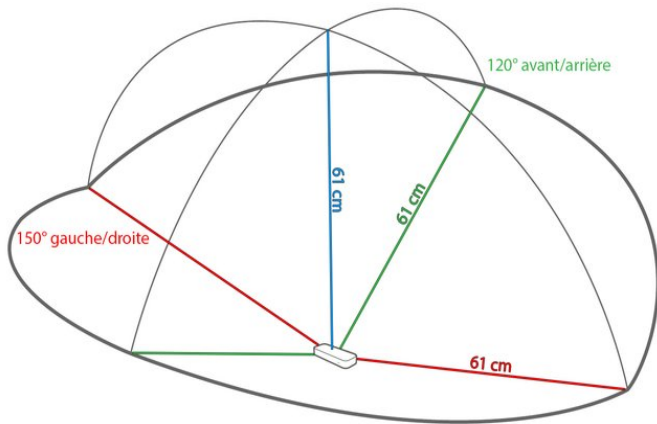


Figure 3. LMC sensing range

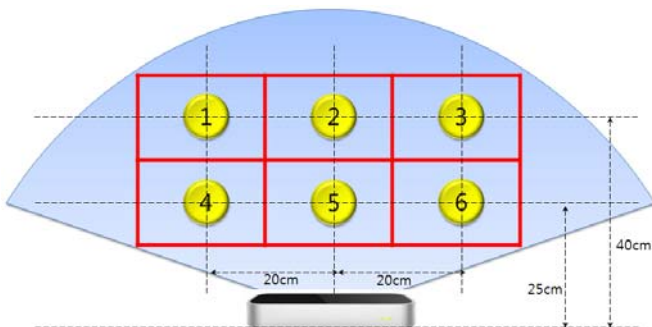


Figure 4. LMC Sensing range

4.2 Implementation of LMC-based hand clapping recognizer

For the experiments of hand clapping recognition, a simple system with the proposed LMC-based hand clapping recognizer has been developed. The output of this system is just a simple window like Fig. 5~7 and it can show the recognition result of current human hand action.

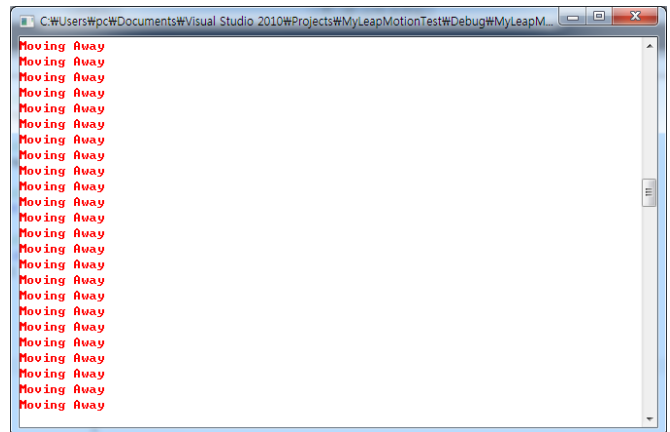


Figure 5. Output of LMC-based hand clapping recognizer when two hands are going away each other, i.e., 'Moving Away' case

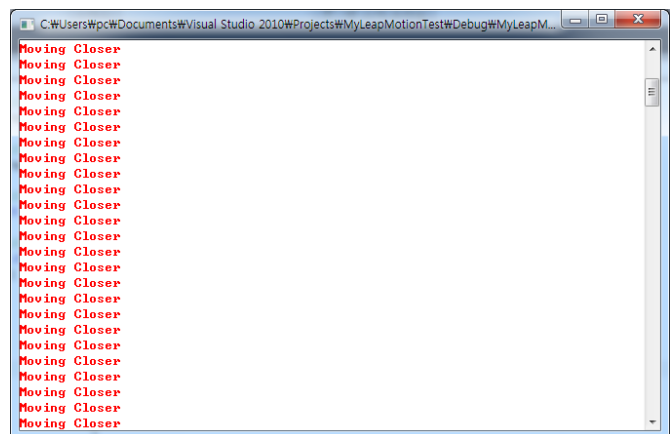


Figure 6. The output of LMC-based hand clapping recognizer when two hands are going closer each other, i.e., 'Moving Closer' case

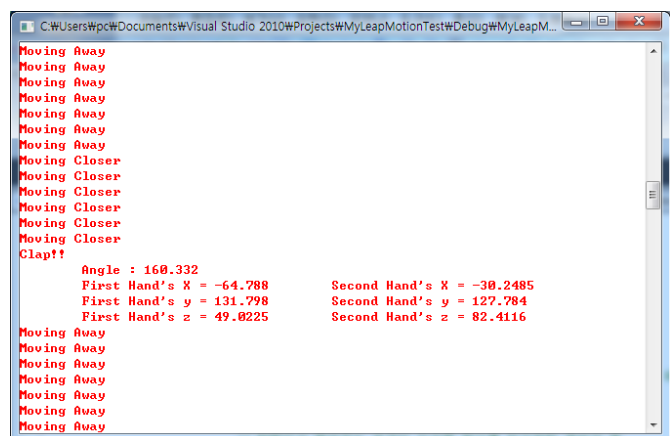


Figure 7. The output of LMC-based hand clapping recognizer when two hands are clapped each other, i.e., 'Clap' case

4.3 Experimental Results

The below tables are showing the experimental result of each test performed by each subject at different hand clapping sub-region. Here, \circ means that hand clapping is correctly recognized once at the clapping moment. Δ means that hand clapping is multiply recognized during one clapping action. And, \times means that hand clapping is not recognized during the whole clapping action.

4.3.1 Experiment I: in case of $Th_{angle}= 180*0.2$ (deg), $Th_{dist}= 40$ (pt), $Th_{wait}=1$ (sec)

Table 1. The result of experiment I with subject1 (Male)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
R1	\circ	\circ	\circ	\times	\circ	\times	\times	\circ	\circ	\circ
R2	\circ	\times	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ
R3	\times	\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ	\circ
R4	\circ	\circ	\circ	\circ	\circ	\circ	\circ	\circ	\times	\circ
R5	\circ	\circ	\times	\circ	\circ	\circ	\circ	\circ	\circ	\times
R6	\circ	\times	\times	\times	\circ	\circ	\circ	\times	\circ	\times
T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	SUM
\circ	\times	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	15
\circ	\times	\circ	\circ	\circ	\times	\circ	\circ	\circ	\circ	16
\circ	\circ	\times	\circ	\circ	\circ	\times	\circ	\circ	\circ	16
\circ	\times	\circ	\circ	\circ	\circ	\circ	\circ	\circ	\circ	18
\circ	\times	\circ	\circ	\circ	\circ	\circ	\times	\circ	\times	15
\times	\circ	\circ	\circ	\times	\circ	\times	\circ	\circ	\circ	12

Table 2. The result of experiment I with subject2 (Female)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
R1	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ	\times
R2	\circ	\circ	\circ	\circ	\times	\times	\circ	\circ	\times	\circ
R3	\times	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ

R4	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ	\circ
R5	\circ	\times	\times	\circ	\circ	\times	\circ	\times	\circ	\circ
R6	\circ	\circ	\times	\times	\circ	\circ	\times	\circ	\circ	\times
T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	SUM
\circ	\times	\circ	\circ	\circ	\times	\circ	\times	\circ	\circ	15
\circ	\circ	\circ	\times	\circ	\circ	\circ	\circ	\times	\circ	16
\circ	\times	\times	\circ	\circ	\circ	\circ	\circ	\circ	\circ	16
\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ	\circ	\circ	18
\times	\circ	\circ	\circ	\times	\times	\circ	\circ	\times	\circ	12
\circ	\circ	\times	\times	\circ	\circ	\times	\times	\circ	\circ	12

Table 3. The result of experiment I with subject3 (Male)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
R1	\circ	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ
R2	\circ	\circ	\times	\circ	\circ	\circ	\circ	\times	\circ	\circ
R3	\circ	\circ	\circ	\times	\circ	\circ	\circ	\circ	\circ	\circ
R4	\circ	\times	\circ	\times	\circ	\circ	\circ	\circ	\circ	\circ
R5	\circ	\circ	\times	\circ	\circ	\circ	\times	\circ	\circ	\circ
R6	\circ	\circ	\circ	\circ	\circ	\circ	\circ	\circ	\circ	\circ
T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	SUM
\times	\circ	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	17
\circ	\circ	\circ	\times	\circ	\circ	\times	\circ	\circ	\times	15
\circ	\circ	\times	\circ	\circ	\circ	\circ	\circ	\circ	\circ	18
\circ	\times	\circ	\circ	\circ	\circ	\times	\circ	\circ	\circ	16
\circ	\circ	\circ	\circ	\times	\times	\circ	\circ	\circ	\times	15
\circ	\circ	\circ	\circ	\circ	\circ	\circ	\times	\circ	\circ	19

4.3.2 Experiment II: in case of $Th_{angle}=180*0.3$ (deg), $Th_{dist}=50$ (pt), $Th_{wait}=1$ (sec)

Table 4. The result of experiment II with subject1 (Male)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
R1	○	×	○	○	○	○	×	○	○	×
R2	○	○	○	○	△	○	○	○	○	○
R3	○	○	×	○	×	○	○	○	○	○
R4	○	○	○	○	○	○	○	×	○	○
R5	○	○	○	○	○	×	○	○	○	○
R6	○	×	○	○	×	○	×	○	○	×

T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	SUM
○	○	○	○	×	○	×	○	○	○	15
○	×	○	○	○	○	○	○	×	○	17
○	×	○	○	○	○	○	×	○	○	16
○	○	○	○	○	×	○	○	○	○	18
○	×	×	○	○	○	○	○	○	○	17
○	○	×	○	○	×	×	○	○	×	12

Table 5. The result of experiment II with subject2 (Female)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
R1	○	○	×	○	○	○	○	○	○	○
R2	○	○	○	○	○	○	○	×	○	○
R3	○	○	○	×	○	○	○	○	○	×
R4	○	○	○	×	○	○	○	○	○	○
R5	○	○	○	○	○	○	○	○	○	○
R6	○	○	○	○	○	○	○	○	○	○

T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	SUM
○	○	○	○	○	×	○	○	○	○	18

○	×	○	○	○	○	○	○	○	○	18
○	○	×	○	○	○	×	○	○	○	16
○	○	×	△	○	○	○	×	○	○	16
○	○	○	○	○	○	○	○	○	○	20
○	○	○	○	○	○	○	○	○	○	20

Table 6. The result of experiment II with subject3 (Male)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
R1	○	○	○	×	○	○	○	○	○	○
R2	○	○	○	○	○	○	○	○	○	○
R3	○	○	○	○	×	○	○	△	○	○
R4	○	○	○	○	○	○	○	○	○	○
R5	×	○	○	○	×	○	○	○	○	○
R6	○	○	×	○	○	○	×	○	○	×

T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	SUM
○	×	○	○	○	○	○	○	○	○	18
○	○	○	○	○	×	○	○	○	○	19
○	○	○	○	○	○	×	○	○	○	17
○	○	○	○	○	○	○	○	○	○	20
○	○	×	○	×	○	○	○	○	○	17
○	○	○	○	○	○	○	○	△	○	16

4.4 Discussion

The recognition rate of experiment A and experiment B are 78.08% and 86.11%, respectively. Even though the recognition rate of experiment B is higher than that of the experiment A, we can also experience more multiple recognition errors with experiment B by releasing the threshold values. Therefore, some more sophisticated method for hand clapping recognition is needed to solve this problem.

5 Concluding Remarks

In this paper, we proposed a method of recognizing human hand clapping action based on the Leap Motion Controller. By using the prior knowledge about basic hand clapping action sequence, LMC-based hand clapping recognizer has been preliminarily developed to show 86.11% accuracy with the three different test users and 240 tests per each user. But, the error of multiple recognition was also increased as the threshold was reduced and its recognition accuracy is increased.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0025247)

6 References

- [1] Yvonne Rogers, Helen Sharp, and Jenny Preece, *Interaction Design: Beyond Human-Computer Interaction*, 3rd., John Wiley & Sons Ltd, 2011
- [2] F A. Weiss, D. Hirshberg, and M. J. Black, "Home 3D body scans from noisy image and range data," 2011 IEEE International Conference on Computer Vision (ICCV), 2011
- [3] Frank Weichert, Daniel Bachmann, Bartholomäus Rudak, and Benis Fisseler, "Analysis of the Accuracy and Robustness of the Leap Motion Controller," *Sensors*, Vol. 13, No.5, pp. 6380-6393, May 2013
- [4] www.leapmotion.com

RAW Video Database of Mobile Video Quality Prediction (MVQP)

Fahad Al Qurashi, Hamad Almohamedh, Ivica Kostanic

Department of Electrical and Computer Engineering

Florida Institute of Technology

Melbourne, Florida, USA

falqurashi2008@my.fit.edu, halmoham@my.fit.edu, kostanic@fit.edu

Abstract - This paper presents the Mobile Video Quality Prediction (MVQP) project. A new QoE tool solution, named MVQP, relies on a non-reference QoE measuring tool. The project aims at developing a tool for the assessment of video streaming quality on mobile phones devices. MVQP project consist of two phases. Phase entails development of a raw video database that contains forty representative 4K raw videos. This video database is accessible free of cost for researchers on MVQP website for the research purpose. The study ends with an evaluation of live video streaming as a part of phase one of MVQP project.

Keywords: MVQP, Video quality measurements, RAW video database.

1 Introduction

Wireless cellular networks are expending both in coverage and capabilities. New 3G and 4G-LTE based radio access offers significant throughput capabilities for each user. This led to a revolutionary increase in mobile-video streaming amongst a large number of users [1]. The increase of data traffic over cellular networks is caused predominantly by video-streaming applications. These applications cause a fluctuation on the data rates, which in turn influence the quality of video streaming service itself [2].

The assessment of the quality of video-streaming is, by no measure, an easy task. This assessment faces various challenges like limited computation powers of the mobile device, distorted videos being found, packet loss, delays and other characteristics and troubles posed by the wireless networks [3].

The Quality of Experience (QoE) measurement of the user is the perceived experience of the service provided by the network. In most networks, the quality of service (QoS) measurements are used to manage the network traffic in a cost-effective manner. QoE guarantees a good relationship between networks and end users [4].

2 MVQP Project

Video streaming over mobile platforms is a very popular data communication service. It is used by all age groups and for various purposes ranging from education to entertainment and infotainment. Network customers are interested in all types of streaming video services. The quality of the video streaming matters whether people are watching a video clip from a movie or a clip from a family occasion having sentimental values. The MVQP project aims to deliver a new scheme to predict the quality of the video streaming using both 3G and 4G-LTE based cellular networks. One finds that calculation and estimation of the quality of the video streaming over the cell phones devices is an issue that has not been satisfactory addressed. At the moment not many solutions have been proposed. Partially, this may be attributed to computational, size and other inherent limitations of the mobile devices.

MVQP project will be able to recognize the quality of the streaming video for a network implementing 3G/LET air interface and for connections utilized by the smart mobile devices.

3 RAW video database generation

The raw video database test bed (RAW) created for MVQP project is a large database of representative video recordings. It consist of various types of shots of movie clips and of various motions. The video database will serve as a good source of videos for the researchers and students alike [9].

3.1 RAW video camera specs

The model of the camera used for the video collection for the RAW database is Sony PMW-F5 CineAlta Digital Cinema Camera 4K. This camera model comes with a super 35mm image sensor. AXS-R5 option recorder was used, and to do so it was mounted behind the camera. An additional high speed flash storage of 512 GB was used in order to work with the 4K RAW options. Choosing this camera model proved to be the right decision as the results were commendable and every little detail was well captured and preserved with the high quality sensor. This model that

has an option to record 16 bit linear capture was used to preserve the details of the image in terms of tones that even the naked human eye would not be able to differentiate.

3.2 Shots locations

Most of the locations have been selected on Florida Institute of Technology (FIT) campus – Melbourne, FL and some of them between Downtown Melbourne and Melbourne beach.

3.3 Type of video shots

3.3.1 Pan

A panoramic view is the one in which the focus of the image remains constant vertically and the camera is moved only horizontally from left to right or vice versa.

3.3.2 Tilt

When a camera moves either upwards or downwards from one fixed initial location, it is called as a tilt.

3.3.3 Tracking

This kind of shot is the following sort. In tracking, the camera actually follows the moving subject to whatever direction it takes in order to not to let the subject out of the view.

3.3.4 Zooming

This kind of shot offers no actual movement of the camera itself. But it works on the principals of physics using the change in focal length of the lens resulting in an impression that the camera actually moved closer or far from the subject.

3.4 Videos names and descriptions

The lists below are a short description for twenty eight of RAW video out of forty in MVQP database, and the snapshot of each of them are shown in Figure 1.

3.4.1 Soccer game (sg)

Shot on a campus on a sunny afternoon. Players are showing diverse contrasts and colors along with complex motions. The camera is tracking the players both sides horizontally.

3.4.2 Sport car (sc)

Shot on the road on a sunny afternoon. The sport car exhibits fast motion with blooming trees on the side of the road. The camera tracks the car from right to left.

3.4.3 Tree (tr)

Shot on campus on a sunny afternoon. Many small leaves are visible, moving slowly. The camera was fixed.

3.4.4 Building (bu)

Shot on campus on a sunny afternoon. The building is surrounded with a large number of trees all around in which leaves are visible moving slowly. The camera pans across the screen from left to right.

3.4.5 Lawn service (ls)

Shot on campus on a sunny morning. A man is providing lawn services by making use of a lawn machine. The camera tracks the machine from left to right.

3.4.6 Pedestrian (pe)

Shot on campus on a sunny morning. Some students are entering while others are leaving. The camera was fixed.

3.4.7 Students at library (sl)

Shot on main campus library on a morning indoors. Students are sitting near a bookrack and are having discussion on a particular topic. The fixed camera zooms out.

3.4.8 Garden (ga)

Shot in a garden on a cloudy morning. There are light color contrasts and slow motion of tree leaves. The camera tilts the trees from bottom to top with a reflection of the cloudy sky.

3.4.9 Turtle (tu)

Shot at a lake on a cloudy morning. The slow movement of turtle within the water has presented a fascinating scene. The camera was slowly tracking the turtle.

3.4.10 Waterfall (wf)

Shot at a small lake on a cloudy morning. There is a fast and steady movement of water flow, which has created a splash. The camera was fixed.

3.4.11 Spider (sp)

Shot on a campus garden on a cloudy morning. The spider exhibits fast motion to complete web. A water flow beneath a spider is presenting complex textures. The camera was fixed.

3.4.12 Large building (lb)

Shot on campus on a sunny morning. Many small leaves are visible, moving slowly in different directions. The camera was moving from the bottom of the building towards the blue sky diagonally.

3.4.13 Presentation (pr)

Shot in a classroom on a morning indoors. Professor is explaining a diagram related to specific topic, displayed on multimedia. The camera was fixed.

3.4.14 Lecture (le)

Shot in a classroom on a morning indoors. The professor is explaining a concept on white board. The camera was fixed.

3.4.15 Tennis training (tt)

Shot in a tennis field on campus on a sunny morning. Two players show diverse color contrasts and complex motions. The camera tracks the players.

3.4.16 Fountain (fo)

Shot near a car parking on campus on a sunny afternoon. The continuous flow of the water is presenting a motion of waterfall. The camera was fixed.

3.4.17 Swimming pool (sw)

Shot on campus at a swimming pool on a sunny afternoon. The water looms towards the camera when the individual made a jump. The camera was fixed.

3.4.18 Fencing activity (fa)

Shot at a fencing area on campus. Players with fencing dress are moving slowly forward and backward. The camera tracks the action.

3.4.19 Melbourne downtown (md)

Shot from the top of the roof on a cloudy afternoon. The entire area is comprised of tall buildings and trees and various cars are moving on the road. The camera pans from right to left.

3.4.20 Bridge (br)

Shot across the bridge in the city of Melbourne on a cloudy afternoon. Various cars are moving on the bridge while water waves are moving slowly downstream. The camera was fixed.

3.4.21 Birds (bi)

Shot near Melbourne beach on a cloudy afternoon. The birds are showing a random and fast movement. The camera was moving slowly with the direction of birds.

3.4.22 Melbourne beach 3 (mb3)

Shot in the city of Melbourne beach on a cloudy afternoon. A girl is moving within the fast waves of water on beach. The camera was fixed.

3.4.23 Playground (pl)

Shot in a park on a sunny afternoon. Children with colorful dresses are enjoying themselves on slides full with bright and fascinating colors. The camera was fixed.

3.4.24 Basketball (ba)

Shot in a park basketball field on a sunny afternoon. Children are getting trained for basketball and a complex motion is being depicted through their movements. The camera was fixed.

3.4.25 Basketball training 2 (bt2)

Shot on campus on an afternoon indoors. Different ratios of light are shown with the movement of players. The camera was showing a steady movement.

3.4.26 House (ho)

Shot in a residential area during nighttime. The garden in front of house is showing a soothing effect and objects around it are all static. The camera was fixed.

3.4.27 Street at night (sn)

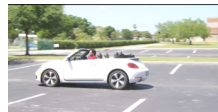
Shot in a street during nighttime. The vehicles are moving at various different speeds and in different directions. The camera was fixed.

3.4.28 Girl talking 2 (gt2)

Shot in a studio on an afternoon indoors. With an extreme close up the girl who is thinking of what to speak on a certain topic. The camera was fixed.



(sg)



(sc)



(tr)



(bu)



(ls)



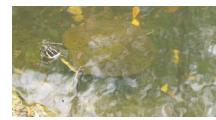
(pe)



(sl)



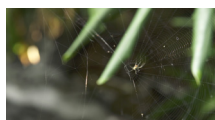
(ga)



(tu)



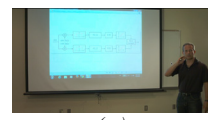
(wf)



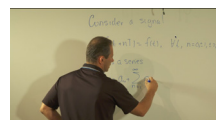
(sp)



(lb)



(pr)



(le)



(tt)



(fo)



(sw)



(fa)



(md)



(br)

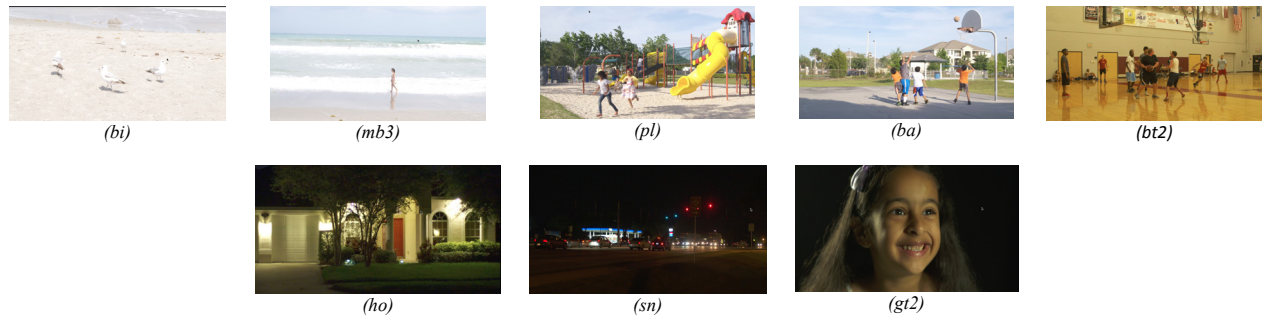


Figure 1. RAW Videos snapshots

4 MVQP website generation

In support for the MVQP development a website has been created. It is up and running and can be accessed at [9]. The website for now has forty RAW YUV videos and allows free access to help the researchers with their studies. The videos captured are of different kinds both in terms of motion and camera movements. Sony F5 camera was used to record 4K RAW videos which were then sampled down to YUV 2K by using a combination of Sony RAW software and VirtualDub which is an open source technique [8].

5 Quality affecting factors

There are some limitations with the re-production of mobile-videos. These mainly come from display size limitations and the processing capabilities of the end terminals. While low bit-rates, frame rates and smaller screen sizes pose more critical issues, spatial resolutions and the quality of frames are also some of the pressing factors that affect video-quality and the quality of the mobile networks [6]. During the transmission of the coded video, many factors can affect the quality of the video that results in distortion, noise and other forms of quality loss at the receiving end.

For this study, a few of the factors affecting the quality of the video streaming are selected to focus on. These factors are the packet loss with some Radio Frequency (RF) factors to determine the video quality over the mobile phones devices over "CDMA2000/ EVDO-Rev A" on Sprint PCS networks.

5.1 Packet loss

Packet loss is the most important factor that affecting the quality of receiving videos. Packet loss defined as a rate at which transmitted packets do not reach at their destination [1].

5.2 Received signal code power (RSCP)

Received signal code power (RSCP) is the measure of power at the receiver end pertaining to a specific physical communication channel. It signifies the strength of the signal [7].

5.3 Interference metric (E_c/I_o)

(E_c) is the received pilot energy, (I_o) is the total power spectral density or alternately the total received energy. Pilot strength is the E_c to I_o ratio that is expressed in dB [5].

6 Live video streaming distortion and RF measurements methodologies

Several videos from the MVQP database are selected to work as a test bed for the experiment. Different videos were selected to keep the test bed as diverse as it could possibly be and to ensure the results were optimal. The factors kept under consideration during the experiment are the Packet loss, RSCP, and (E_c/I_o). "CDMA2000/ EVDO-Rev A" on Sprint PCS was used to stream the videos and to measure the impact of the above mentioned distortion factors. The results revealed a relationship between the factors, which is shown in part 7.

7 Results

As a preliminary illustration two videos have been selected. The first one is a pedestrian clip, and the second one is a bridge clip. To reference the clips, they are named as (pe) and (br) respectively. One of them is a fast motion and the other one is a slow motion. Sprint PCS mobile networks is used for the live video streaming experiments.

The graphs in Figures 2 and 3 shows the results of the analysis. It is evident from the graphs for both the videos that the packet loss is inevitable if the signal is either low or high with the presence of interference.

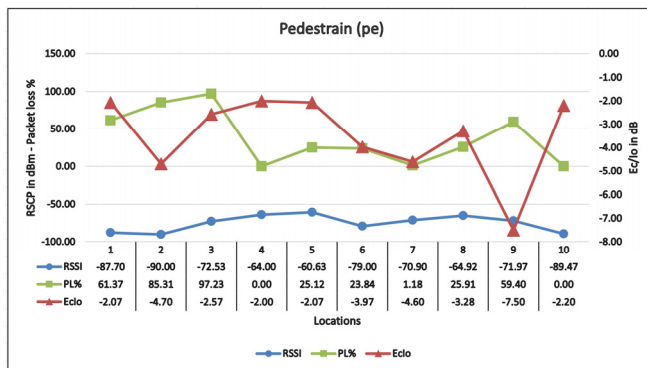


Figure 2. Pedestrian (pe) video streaming analysis.

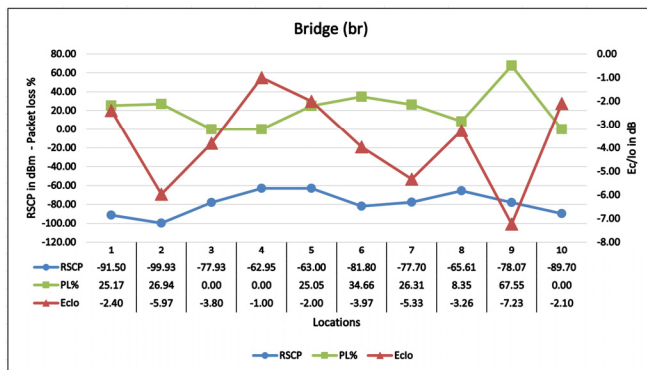


Figure 3. Bridge (br) video streaming analysis.

8 Conclusion

This study proposed an introduction of MVQP that will be the tool of prediction of the quality of video streaming over the mobile phones devices. The study was backed up with the experimental results that showed a relationship between the considered distortion factors. In [10], [11], and [12] we extended this study with another experiments that contained the in depth analysis of the experiments for 3G/4G-LTE video quality streaming measurements.

9 References

- [1] Chi-Wen Lo ; Dept. of Electr. Eng., Nat. Tsing Hua Univ., Hsinchu, Taiwan ; Lin, Chia-Wen ; Yung-Chang Chen ; Jen-Yu Yu. "A packet loss estimation model and its application to reliable mesh-based P2P video streaming"; Multimedia and Expo (ICME), 2011 IEEE International Conference on July 2011.
- [2] Ayaskant Rath, Sanjay Goyal, and Shivendra Panwar; "Streamloading: Low Cost High Quality Video Streaming for Mobile Users"; Department of Electrical and Computer Engineering, Polytechnic Institute of New York University.
- [3] An (Jack) Chan, Amit Pande, Eilwoo Baik, Prasant Mohapatra. "Temporal Quality Assessment for Mobile

Videos"; Mobicom '12: Proceedings of the 18th annual international conference on Mobile computing and networking; August 2012.

[4] Meral Shirazipour, Gregory Charlot, Geoffrey Lefebvre, Suresh Krishnan, Samuel Pierre. "ConEx based QoE feedback to enhance QoE"; CSWS '12: Proceedings of the 2012 ACM workshop on Capacity sharing; December 2012

[5] Rockstar Bidco Lp. "System and Method for Ec/Io Access Screening in a CDMA Network" in Patent Application Approval Process Politics & Government Week (Jul 25, 2013): 2666.

[6] Satu Jumisko-Pyykkö. "Evaluation of subjective video quality of mobile devices"; MULTIMEDIA '05 Proceedings of the 13th annual ACM international conference on Multimedia Pages 535 - 538

[7] Zhenqiang Wang ; Yi-Sheng Zhu ; Nan Liu Electronic and Mechanical Engineering and Information Technology (EMEIT). " An improved mobile location approach based on RSCP difference "; 2011 International Conference on Volume: 6 Digital Object Identifier: 10.1109/EMEIT.2011.6023675 Publication Year: 2011 , Page(s): 2765 – 2768

[8] Virtualdub open source Internet: <http://www.virtualdub.org/index.html>

[9] Ivica Kostanic, Fahad Al Qurashi, Hamad Almohamedh, "MVQP Project "; Internet: <http://research.fit.edu/wice/mvqp.php>, Mar. 2014

[10] Fahad Alqurashi, Hamad Almohamedh, Ivica Kostanic. "Subjective Video Streaming Quality Evaluation in 3G Cellular Networks"; The 2014 International Conference on Image Processing, Computer Vision, and Pattern Recognition; (in-press), 2014.

[11] Hamad Almohamedh, Fahad Alqurashi, Ivica Kostanic. "Mobile Videos Quality Measurements over Long Term Evolution (LTE) Network"; The 2014 International Conference on Image Processing, Computer Vision, and Pattern Recognition"; (in-press), 2014.

[12] Hamad Almohamedh, Fahad Alqurashi, Ivica Kostanic. "Subjective Assessment of Mobile Videos Quality for Long Term Evolution (LTE) Cellular Networks"; World Congress on Engineering and Computer Science (WCECS 2014); (in-press), 2014.

Geometrical Analysis Based Text Localization Method

Samabia Tehsin¹, and Asif Masood²

¹CS Department, MCS, National University of Sciences and Technology, Islamabad, Pakistan

²MCS, National University of Sciences and Technology, Islamabad, Pakistan

Abstract - Textual information embedded in multimedia can provide a vital tool for indexing and retrieval. Text extraction process has many inherent problems due to the variation in font sizes, color, backgrounds and resolutions. Localization of detected text is the most challenging phase. Text extraction results are highly dependent upon this phase. This paper focuses on the text detection and localization because of their very fundamental importance. A text detection and localization methodology along with the geometrical analysis of text objects is presented.

Keywords: Text extraction, Caption text, Geometric analysis, ICDAR 2011.

1 Introduction

With the dramatic increase in multimedia data, escalating trend of internet, and amplifying use of image/video capturing devices; content based indexing and text extraction is gaining more and more importance in research community. Embedded text in images/videos can be very instrumental for data retrieval as visual texts of multimedia data often impart knowledge about news headings, title of movie, brands of products, scores of a sports contest, date and time of events etc. Such information can be influential for understanding and retrieval of images or videos.

Text extraction process includes text detection, localization, tracking, binarization, and recognition. The initial three modules are very significant to attain promising results. Aim of text detection and localization phase is to mark a bounding box around all the text objects appearing in the image. Text embedded in images and videos may be categorized in two groups, namely, caption text and scene text. Caption text is imposed over the image/video at the editing stage e.g. score of match and name of the speaker. It is also known as artificial text or superimposed text. Whereas scene text is an actual part of the scene i.e. name of the product during commercial break, sign board, name plate and text appearing on dresses or product, etc.

In the last decade, a lot of techniques for text extraction have been reported in the literature. Section 2 focuses on the text detection and localization techniques, section 3 provide the proposed methodology in detail, section 4 presents the evaluation of the presented work and section 5 concludes the research paper.

2 Literature Review

A variety of techniques for text extraction have appeared in recent past [1-6]. Comprehensive surveys can be traced explicitly in [7-9]. These techniques can be categorized into two types mainly with reference to the utilized text features i.e. region based and texture based [10]. Texture based methods pertain to textural properties of the text, distinguishing it from the background. The techniques mostly use Gabor filters, Wavelet, FFT, spatial variance, etc. These methods further use machine learning techniques such as SVM, MLP and adaBoost [11-14]. These techniques work in the top down fashion by first extracting the texture features and then finding the text regions.

Region based approach exploits different region properties to extract text objects. This approach makes use of the fact that there is sufficient difference between the text color and its immediate background. Color features, edge features, and connected component methods are often used in this approach [15-17]. These techniques typically work in the bottom up fashion by first segmenting the small regions and then grouping the potential text regions. Texture based techniques usually give better results in complex backgrounds than region based techniques but have computationally very heavy hence not suitable for retrieval systems for hefty databases. So there is a need to improve the detection results of region based techniques, to be used for retrieval and indexing of large multimedia data.

3 Proposed Methodology

Region based techniques typically work in the bottom up fashion by initially segmenting the small regions and lately grouping the potential text regions. Proposed methodology extracts the text in three modules. (1) Text candidate extraction, (2) Merging and grouping of text candidates (3) Differentiating between text and non text objects.

3.1 Segmentation

Given an input image, goal of this process is to repeatedly partition the image into a small number of regions that are consistent in color and composition. This process is divided into two sub-processes. The first one defines the segmentation of image into k regions for specified value of k . The second sub-process describes the mechanism to automatically select the value of k , the number of regions.

Joining the two sub-processes, gives the solution to the image segmentation for text detection process.

Color based k-means clustering is chosen for the first sub-process. K-means clustering is the promising and thoroughly researched image segmentation methodology. There exists sharp contrast between the text and its background. Because of this contrasting nature of text, pixels of text object and its background generally assigned to different clusters. Due to the low resolution of web images, contrast enhancement is applied before the clustering process. The objective of contrast stretching is to get sharper contrasts for segmentation.

$$h = 255 \frac{[\omega_c(\delta_g) - \omega_c(\delta_{gmin})]}{[\omega_c(\delta_{gmax}) - \omega_c(\delta_{gmin})]} \quad (1)$$

Where δ_g is the sigmoid function for $C \times C$ sliding window and is defined as

$$\omega_c(\delta_g) = \left[1 + \exp\left(\frac{m_c - g}{\sigma_c}\right) \right] \quad (2)$$

δ_{gmin} and δ_{gmax} are the minimum and maximum intensity values for the gray scale input image. m_c and σ_c are the mean and variance of grayscale intensity values of the given window. Contrast stretching process is applied on the three channels of the colored image independently.

Suppose $I: \Omega \rightarrow \mathbb{R}$ is the input image to be segmented, where $\Omega \subset \mathbb{R}^3$ is the domain of I . For given image I , the goal of image segmentation is to partition its domain into K distinct regions. Precisely, segmentation defines the set of disjoint regions $\{\Omega_i\}_{i=1}^K$, such that $\Omega = \cup_{i=1}^K \Omega_i$. Here, Ω_i presents the i^{th} region of the image I .

For given set of observations $S = \{\bar{s}_1, \bar{s}_2, \dots, \bar{s}_n\}$, k-means partition the n observations into k sets $\{\Omega_i\}_{i=1}^K$, for specified value of k .

$$\ddot{X} = \sum_{i=1}^k \sum_{j=1}^n \ddot{\mu}_{ij} \|\bar{s}_j - \bar{\zeta}_i\|^2 \quad (3)$$

where $\zeta = \{\bar{\zeta}_1, \bar{\zeta}_2, \dots, \bar{\zeta}_k\}$ are the centers for the k clusters.

$$\bar{\zeta}_i = \frac{\sum_{j=1}^n \ddot{\mu}_{ij} \bar{s}_j}{\sum_{j=1}^n \ddot{\mu}_{ij}} \quad (4)$$

Partition matrix $\Upsilon = \{\ddot{\mu}_{ij}\}$ can be defined as

$$\ddot{\mu}_{ij} = \begin{cases} 1 & \text{if } \bar{s}_j \in \Omega_i \\ 0 & \text{otherwise} \end{cases}$$

The centers and partition matrices are updated until the solution converges [18]. Different methods to choose the

initial cluster centroid positions, sometimes known as seeds, are presented in the literature [19, 20]. In the projected system, K points are chosen uniformly at random from the domain Ω [21]. This seeding method capitulate substantial enhancement in the ultimate error of k-means. With this initial selection method, the k-means algorithm converges very swiftly and thus reduces the computational cost of the segmentation process.

Determining the optimal number of clusters is very vital task for the clustering process. Performance of the k-means clustering is highly dependent on the choice of K . The right selection of K is usually indefinite. Increasing K may lead to over segmentation of the image, and decreasing the value of K may ends up with under segmentation issues. Intuitively then, there is a need to determine the optimal choice of K that can give the desirable segmentation results. There are numerous classes of techniques for building this decision [22-24]. Three factors are used for determining the number of clusters for the input image. Empirical study proves that K having value 2 or 3 gives the best results. Number of clusters, are determined using three gray level co-occurrence matrices (GLCM) features applied on L component of input image I . These features are energy (\mathfrak{E}), entropy (\mathfrak{P}) and contrast (\mathfrak{C}) and can be defined as

$$\mathfrak{E} = \max_{\theta} \left(\sum_i \sum_j P_d^2(i, j) \right) \quad (5)$$

$$\mathfrak{P} = \max_{\theta} \left(- \sum_i \sum_j P_d(i, j) \log P_d(i, j) \right) \quad (6)$$

$$\mathfrak{C} = \max_{\theta} \left(\sum_i \sum_j (i - j)^2 P_d(i, j) \right) \quad (7)$$

A GLCM element $P_{\theta,d}(i, j)$ is the joint probability of the pixel pairs with gray levels i and j in a given direction θ , having d distance between them. Here $d=1$ and $\theta = \{0, 90, 180, 270\}$.

3.2 Grouping of text characters

The merging of character candidates rely on number of similarity factors. Five features are extracted to describe the similarity of two character candidates. These features are color, height, position, distance and stroke width.

$$\mathbf{Color:} \quad \sqrt{(C_i^1 - C_j^1)^2 + (C_i^2 - C_j^2)^2 + (C_i^3 - C_j^3)^2} \quad (8)$$

$$\mathbf{Height:} \quad \Delta H_{i,j} = \frac{|H_i - H_j|}{H_i} \quad (9)$$

$$\mathbf{Position:} \quad \Delta Pos_{i,j} = \frac{|Pos_i - Pos_j|}{H_i} \quad (10)$$

$$\text{Distance: } \Delta d = \frac{\min(|x_i(1) - x_j(2)|, |x_j(1) - x_i(2)|)}{H_i} \quad (11)$$

$$\text{Stroke width: } \frac{|S_i - S_j|}{\max(S_i, S_j)} \quad (12)$$

C_i^1, C_i^2 and C_i^3 be the average three channel color value of i ; H_i and H_j are the heights, Pos_i and Pos_j are the bottom coordinates of bounding boxes, S_i and S_j are the stroke widths of i^{th} and j^{th} objects respectively; $x_n(1)$ and $x_n(2)$ are the left and right coordinates of bounding box of n^{th} object.

3.3 Text candidate selection

The classification feature vector is generated by the investigation of the geometric structure of text content. Dissection of textual data shows that geometry of text objects is quite different from the non-text ones. Two very effectual geometric features are introduced in this section. Both are based on the geometric characteristics of text.

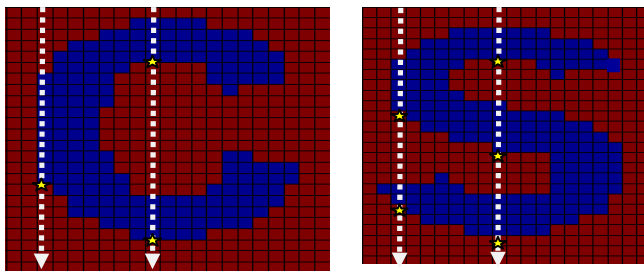


Figure 1: Perpendicular transition

It is observed in the text elements that these are having transition in their patterns. Transition here is used as the alteration of the foreground pixel to background pixel. In figure 1, 'star' is representing the perpendicular transition \square , i.e. transition occurs in particular column of the image. \mathbb{W}_j is the perpendicular transition count of column j . In figure 1, 'C' has $\mathbb{W}_3 = 1$ and $\mathbb{W}_{10} = 2$; where 'S' has $\mathbb{W}_3 = 2$ and $\mathbb{W}_{10} = 3$.

Figure 2 shows maximum and minimum perpendicular transition count for all the capital/small English alphabets and numeric digits. It can be observed that most of the digits have this count between one and three except 'g'. Letter 'g' may have four transitions in some fonts but may not have four in other fonts (g). Dark grey color in Figure 2 represents the possible range of the count for textual data. So the every column of the text object should have the perpendicular transition count between 'one' and 'four'. Mathematically,

$$\mathbb{W}_j = \sum_{i=2}^H \nabla_{i,j} \quad (13)$$

$$\nabla_{i,j} = \begin{cases} 1 & \text{if } \Omega(i-1, j) \in \text{foreground and } \Omega(i, j) \in \text{background} \\ 0 & \text{otherwise} \end{cases}$$

Here 'H' is the row height of the text candidate Ω .

For text objects the perpendicular transition count for every column must be in the range of 1-4. On the other hand, non-text instances have very low probability of having such distribution. Here is the mathematical description of vertical spread for n^{th} object

$$\underline{w}_n = \frac{\text{length}(\text{find}(0 < \mathbb{W} \leq 4))}{(\hat{W} - \underline{3})} \quad (14)$$

Here, $\mathbb{W} = \{\mathbb{W}_1, \mathbb{W}_2, \dots, \mathbb{W}_{\hat{W}}\}$, \hat{W} is the column width of the object and $\underline{3}$ are the empty columns i.e. the columns with no foreground pixel. Numerator shows the length of the vector having perpendicular transition counts between the specified range. For ideal text object, the feature \underline{w} is exactly 'one'. But in practice this may not achieve the 'ideal' value due to noise, blur and low resolution images. Experimentation has proved this feature as the good option for classification. Figure 3(a) present the box plot for the vertical spread and it is evident from the plot that majority of the text object has its value ≈ 1 .

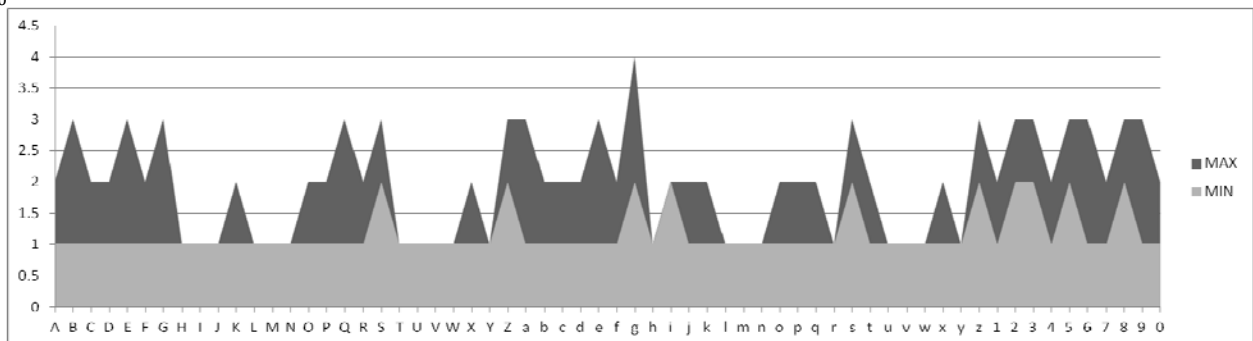


Figure 2: Range of perpendicular transition count for textual data

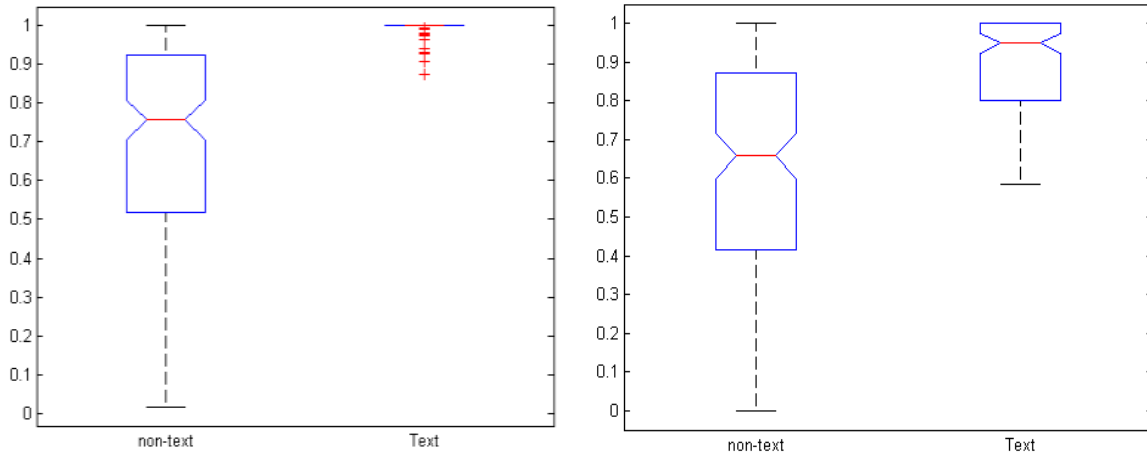


Figure 3: Box plot (a)vertical spread (b) horizontal spread

Horizontal transitions are also studied for classification of text objects. Interesting findings are achieved through the analysis of text content. It is observed through the experimentation that at least one horizontal transition per character is visible in each and every row of the text objects. Horizontal spread $\overline{\mathfrak{M}}$ of nth object can be defined as

$$\overline{\mathfrak{M}}_n = \frac{\text{length}(\text{find}(\mathfrak{M} \geq \eta))}{(H - \overline{\mathfrak{Z}})} \quad (15)$$

Here η is the number of characters, H is the height and $\overline{\mathfrak{Z}}$ are the empty rows of the given text object. $\mathfrak{M} = \{\mathfrak{M}_1, \mathfrak{M}_2, \dots, \mathfrak{M}_H\}$ is the vector of the horizontal transition counts. Mathematically,

$$\mathfrak{M}_j = \sum_{i=2}^{\hat{w}} \Delta_{i,j} \quad (16)$$

$$\Delta_{i,j} = \begin{cases} 1 & \text{if } \Omega(i, j - 1) \in \text{foreground and } \Omega(i, j) \in \text{background} \\ 0 & \text{otherwise} \end{cases}$$

Figure 3(b) shows the box plot for the horizontal spread. This plot presents the potential strength of this feature as text object identifier.

The proposed geometric feature vector has two elements namely; vertical spread and horizontal spread and is defined as $f_n^2 = \{\overline{\mathfrak{v}}_n, \overline{\mathfrak{M}}_n\}$. Combined strength of horizontal and vertical spread can be visualized in Figure 4. The scatter plot explicitly defines the collective potency of geometric feature

vector. Reduced length of feature vector ensures the fast computation in the classification stage.

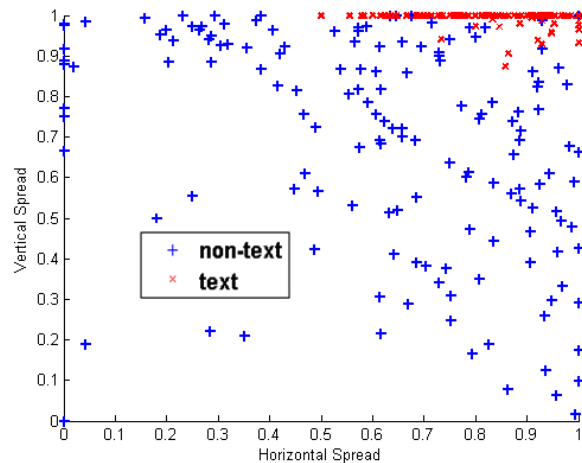


Figure 4: Scatter plot for geometric feature vector

4 Result and analysis

International Conference on Document Analysis and Recognition (ICDAR) is the most renowned Conference on Analysis and Recognition of text in images. Research competition is a popular segment of this conference.

Researchers from all over the world bring their research and participate in various categories of this competition. ICDAR presented compact and versatile datasets for the Robust Reading Competitions. Datasets of ICDAR 2011[25] for Robust Reading Competition, Challenge 1: "Reading Text in Born-Digital Images (Web and Email)", are widely used datasets for benchmarking of born-digital images and hence used for the evaluation of this research.



Figure 5: Results of proposed methodology

Figure 5 shows the results of proposed approach on the images from ICDAR 2011 dataset.

[29]. Zeng et al. [27] and Gonzalez et al. [28] are reported in 2013 and 2012 respectively.

Table 1: Comparison of proposed work on ICDAR 2011 dataset

Method	Recall	Precision	Harmonic Mean
Proposed	78.65	85.35	81.86
Textorter[26]	69.62	85.83	76.88
TH-TextLoc	73.08	80.51	76.62
TDM_IACAS	69.16	84.64	76.12
OTCYMIST	75.91	64.05	69.48
SASA	65.62	67.82	66.7
Text Hunter	57.76	75.52	65.46
Zeng et al. [27]	74.88	85.35	79.78
Gonzalez et al. [28]	70.08	89.23	78.51

Table 1 shows the results of the proposed technique, in comparison with other techniques. Textorter[26], TH-TextLoc, TDM_IACAS, OTCYMIST, SASA and Text Hunter are the participants of the ICDAR 2011 competition

5 CONCLUSION

Content based image and video retrieval has generated immense interest for researchers due to its applicability and utility. Text objects appearing in multimedia content is a vital tool for content based image/video retrieval and indexing. This text presents a much useful semantic knowledge about the contents of the multimedia document. In this paper, a novel methodology is presented for text detection and localization task. The presented research work is evaluated on ICDAR 2011 Robust Reading dataset, and provides excellent results.

6 REFERENCES

1. Minetto, R., Thome, N., Cord, M., Leite, N. J., & Stolfi, J. (2012). T-HOG: An effective gradient-based descriptor for single line text regions. *Pattern Recognition*.
2. Neumann, L., & Matas, J. (2013) On Combining Multiple Segmentations in Scene Text Recognition. *12th*

- International Conference of Document Analysis and Recognition (ICDAR)*
3. Zhao, M., Li, S., & Kwok, J. (2010). Text detection in images using sparse representation with discriminative dictionaries. *Image and Vision Computing*, 28(12), 1590-1599.
 4. Wang, K., & Belongie, S. (2010). Word spotting in the wild. In *Computer Vision–ECCV 2010* (pp. 591-604). Springer Berlin Heidelberg.
 5. Neumann, L., & Matas, J. (2011). A method for text localization and recognition in real-world images. In *Computer Vision–ACCV 2010* (pp. 770-783). Springer Berlin Heidelberg.
 6. Shivakumara, P., Phan, T. Q., & Tan, C. L. (2011). A laplacian approach to multi-oriented text detection in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2), 412-419.
 7. Jung, K., In Kim, K., & K Jain, A. (2004). Text information extraction in images and video: a survey. *Pattern recognition*, 37(5), 977-997.
 8. Liang, J., Doermann, D., & Li, H. (2005). Camera-based analysis of text and documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2-3), 84-104.
 9. Sumathi, C.P., Santhanam, T., & Gayathri G. (2012). A Survey on various approaches of text extraction in images. *International Journal of Computer Science & Engineering Survey (IJCSES)*, 3(4).
 10. Lienhart, R. (2003). Video OCR: a survey and practitioner's guide. In *Video mining* (pp. 155-183). Springer US.
 11. Li, C., Ding, X. G., & Wu, Y. S. (2006). An Algorithm for Text Location in Images Based on Histogram Features and Ada-boost. *Journal of Image and Graphics*, 3, 003, 325-331
 12. Kim, K. I., Jung, K., & Kim, J. H. (2003). Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12), 1631-1639.
 13. Gllavata, J., Qeli, E., & Freisleben, B. (2006, December). Detecting text in videos using fuzzy clustering ensembles. In *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on* (pp. 283-290). IEEE.
 14. Chen, D., Odobez, J. M., & Bourlard, H. (2004). Text detection and recognition in images and video frames. *Pattern Recognition*, 37(3), 595-608.
 15. Shi, C., Wang, C., Xiao, B., Zhang, Y., & Gao, S. (2012). Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters*.
 16. León Cristóbal, M., Vilaplana Besler, V., Gasull Llampallas, A., & Marqués Acosta, F. (2012). Region-based caption text extraction. *11th International Workshop On Image Analysis For Multimedia Interactive Services (Wiamis)*
 17. Epshtein, B., Ofek, E., & Wexler, Y. (2010, June). Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 2963-2970). IEEE.
 18. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7), 881-892.
 19. Hamerly, G., & Elkan, C. (2002, November). Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 600-607). ACM.
 20. Ostrovsky, R., Rabani, Y., Schulman, L. J. and Swamy, C. (2006). "The Effectiveness of Lloyd-Type Methods for the k-Means Problem". *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. pp. 165–174.
 21. Arthur, D., & Vassilvitskii, S. (2007, January). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.
 22. Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7), research0036.
 23. Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463).
 24. Lleti, R., Ortiz, M. C., Sarabia, L. A., & Sánchez, M. S. (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1), 87-100.
 25. <http://www.cvc.uab.es/icdar2011competition/>
 26. Tehsin S, Masood A, Kausar S, Javed Y., "Text localization and detection method for born-digital images" *IETE J Res* 2013;59:343-9
 27. Zeng, C., Jia, W., & He, X. (2013, May). Text detection in born-digital images using multiple layer images. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 1947-1951). IEEE.
 28. González, A., Bergasa, L. M., Yebes, J. J., & Bronte, S. (2012, November). Text location in complex images. In *Pattern Recognition (ICPR), 2012 21st International Conference on* (pp. 617-620). IEEE.
 29. Karatzas, D., Mestre, S. R., Mas, J., Nourbakhsh, F., & Roy, P. P. (2011, September). ICDAR 2011 Robust reading competition-challenge 1: reading text in born-

digital images (web and email). In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (pp. 1485-1490). IEEE.

A Comparison of Local Descriptors on Cardiac Ultrasound Images

Meng Ma¹ and Xin Yang¹

¹Department of Automation, Shanghai Jiao Tong University, Shanghai, China

Abstract—*In the literature of pattern recognition and computer vision, local descriptors have been widely used in applications such as shape matching and object recognition. Numerous descriptors have been proposed and evaluated, but little work is reported in the area of medical image, especially ultrasonic images. In this paper, we assess the performance of different local descriptors to detect specific objects in cardiac ultrasound image. Ultrasonic images are particular noisy. It is yet to be determined whether the descriptors from general vision problems can still have ideal performance on ultrasound images. We compare several descriptors such as context, histogram, moment invariant, texture and generic Fourier descriptor (GFD), and use recall and 1-precision to evaluate their performance. Experiments show that moment invariant and GFD have higher recall, but high 1-precision as well. Combination of different descriptors are also evaluated and they turn out to be more effective than single descriptors.*

Keywords: cardiac ultrasound descriptor SVM

1. Introduction

Local descriptors have become increasingly popular in applications such as shape matching, image retrieval, object recognition, etc. They prove to be very effective to represent image patches. Generally, descriptors fall into two categories: *global* descriptors and *local* descriptors. Global descriptors take the whole image as input, and capture overall features. Typical applications include image retrieval and image classification. Local descriptors, however, only exploit part of the image to extract feature vectors, thus being able to distinguish details between different parts of an image. Local descriptors can be further classified into two categories: *point-based* and *region-based*. *Point-based* descriptors rely on some key points, making them vulnerable to noise. Typical methods include corners, edges, auto-context, and so on. *Region-based* methods, on the other hand, use all the pixels in the patch to extract feature vectors. Obviously, they are more likely to be noise insensitive, thus more robust.

A lot of work has been done to evaluate descriptors in vision applications, where natural images are dominant. Carneiro and Jepson [16] evaluate the performance of phase-based point descriptor. Zhang and Lu [5] compare various shape representation techniques and come to the conclusion

that Fourier Descriptors, especially GFD, outperform other methods in both accuracy and error rate. Lowe [9] propose the famous SIFT descriptor. Ke and Sukthankar [10] develop PCA-SIFT descriptor and show that it outperforms SIFT descriptor. Mikolajczyk and Schmid [7] compare the performance of 10 different descriptors.

However, little work is done in the evaluation of local descriptors on medical images. In this paper, the performance of five commonly used local features are evaluated on cardiac ultrasound (US) images. Since US images are quite different from those in computer vision, some local descriptors fail to perform well. We also propose two combined methods that incorporate simple descriptors.

2. Method

In this paper, the goal of our work is to assess the capability of various local descriptors to characterize image patches on ultrasound images. Our evaluation is based on the project that aimed to assist doctors' diagnosis using cardiac US images. The primary purpose is to automatically exclude patients who are absolutely normal and leave those who are likely to have heart problems for further diagnosis, thus reducing the workload of doctors. In our experiments, the task is to use pattern recognition to automatically detect mitral valve. We employ several local descriptors in this task and make a comparison of their performance. Our procedure can be mainly divided into four steps: preprocessing, feature extraction, classification and post processing.

2.1 Preprocessing

Cardiac ultrasound image has many advantages such as cheap, fast, no radiation and high dynamic range, making it an important tool in heart disease diagnosis. However, it also suffers from many drawbacks, including limited view field, dependency on skilled operators and particularly noisy image. It is difficult to apply many algorithms with so much noise, therefore, some preprocessing is needed before experiment. Much work has been done to overcome the relatively low image quality. Rocha, Silva and Campilho [2] use a Gaussian filter to remove speckles and smooth the image before applying main algorithm. Aysal and Barner [22] have showed that the noise is actually multiplicative Rayleigh noise for which mode filter is optimal. Since point-based methods are easily corrupted by noise, some measures

must be taken to suppress noise first. Mode filter is computational expensive and even meaningless for continuous signal, Davies [17] introduce the truncated median filter for an approximation. It is based on the fact that for many non-Gaussian distributions, the order of mean, median and mode is the same. If we truncate the distribution, the median will approach the mode. The window size of filter cannot be too small otherwise it would not suppress noise effectively. According to [23] and our experiment, a 7×7 size window would be large enough.

2.2 Feature Extraction

In the following, we will discuss details of feature extraction techniques. Feature vectors are computed from patches using different descriptors. Distinguishability generally depend on describing techniques, but poor selection of patches can have a negative influence.

2.2.1 Patch Size

Patch is the region used to compute descriptors. Though some point-based features only need sparse points around the center to extract feature, there are also a lot of region-based descriptors that take all patch as input. First, the region should be large enough to provide adequate information to distinguish different part, but larger size means more computational complexity. Moreover, large region may have some irrelevant parts included, which offer no help and only serves as noise. Thus, this is a trade-off problem, and the actual region size should depend on image size, resolution, and the descriptor used.

2.2.2 Descriptors

Feature extraction plays a vital role in pattern recognition. But there are no general rules that can tell the good features from the bad ones, so this is an application dependent problem. To accomplish our task, a natural idea is to borrow the good features from some similar field, such as vision or artificial intelligence. However, due to the specialty of medical images, descriptors may fail to perform as well as in original field, so some modification may be needed. In the following, we present the details of the descriptors used in our evaluation. We use five simple descriptors: context, histogram, moment, texture, and GFD.

Context is a point-based descriptor. Tu and Bai[6] have used auto-context model in 3D brain image segmentation for its simple implementation and good performance. They introduce an iterative method that takes advantage of results from previous iteration. Input image is first processed by classifier 1, and then this image, as well as the classification result is processed classifier 2. Then classifier 3, 4, ..., and so on. This method, however, is not used here because want to evaluate the distinguish ability of descriptors itself. The configuration is illustrated in Fig. 1(a), and classification result is shown by (b), (c), and (d).

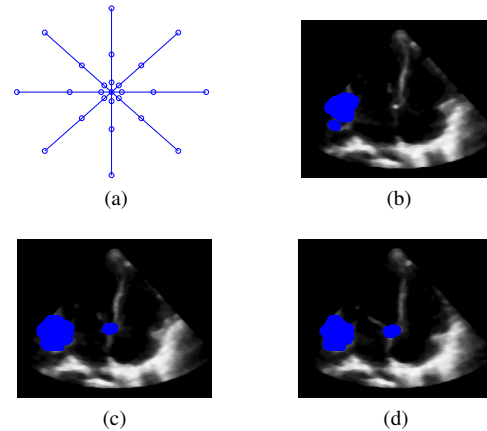


Fig. 1: Context Descriptor

Histogram is a simple yet powerful feature. It is based on 1st order statistic of an image and shows how individual brightness levels are occupied in an image. The histogram of a digital image with L total possible intensity levels in the range $[0, G]$ is defined as the discrete function

$$h(r_k) = n_k \quad (1)$$

where r_k is the k -th intensity level in the interval $[0, G]$ and n_k is the number of pixels in the image whose intensity level is r_k . The classification result is shown in Fig. 2(a), (b), and (c). Clearly, histogram has a strong ability to detect similar points, but it suffers from a high false positive rate. With some modification, that is, combined with context feature, we can greatly improve its performance. The classification result of this combined descriptor is shown in Fig. 2(d), (e), (f).

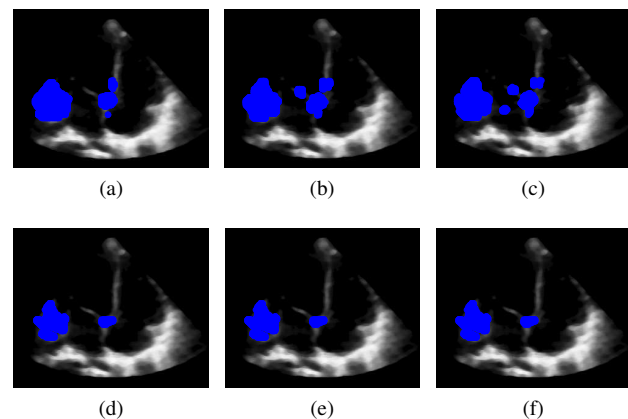


Fig. 2: Histogram Descriptor

Moment is a globe description of shape that combines some low level features, such as area, compactness and irregularity together. Moment are often computed up to the

second order and second degree. For digital images, the 2D-moment of $(p + q)$ degree is defined as

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \quad (2)$$

for $p, q = 0, 1, 2, \dots$ where the summations are over the spatial coordinates. A set of seven 2D moment invariants that are insensitive to translation, scale change, mirroring and rotating can be derived from these equations. Details about moment invariant can be found in [5]. Moment invariants are robust, efficient and easy to compute. However, these moment invariant assumes that images are not sparse. This is usually true for most vision applications, but not necessarily for medical images. In fact, nearly all the images, including MRI, CT and ultrasound, contain large region of darkness, making it difficult to compute moment invariant. Some measure can be taken to adjust these images such as giving the dark pixel a slight positive intensity instead of zero. Such measures may work, but the result is still unsatisfactory.

Texture is an important region-based descriptor that are often used to characterize properties of material, such as smoothness, coarseness and regularity. The commonly used approach to describe texture is statistical operator, which make use of statistical moments as texture description. The n -th central moment can be defined as

$$\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i) \quad (3)$$

where z_i is intensity, $p(z_i)$ is the histogram, L is the number of intensity levels, and m is the average intensity values. Here, we use 5 statistical moment to compute texture, that is, mean (m), standard deviation (σ), third moment (μ_3 , a measure of skewness of the histogram), uniformity ($U = \sum_{i=0}^{L-1} p^2(z_i)$) and entropy ($E = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$). These moments give a good description of image histogram, and carry no information about relative coordinates of pixels. When describing texture merely, this is an important property, but fails to fulfill the requirement to distinguish structure parts, as is illustrated in Figure 3 (a)–(c). There are many false positive misclassification in (c). One way to incorporate position information is by combing texture and context. Now, the combined descriptor have both the ability to characterize texture as well as structural information. And the result Figure 3(d)–(f) show that it outperforms the original ones.

Fourier descriptors, allow us to use Fourier analysis theory that has been proved very effective and being the most popular tools for decades in signal processing. The basic idea of Fourier analysis is a transformation from spatial domain to frequency domain. Spatial domain is the one that reflects spatial structure and can be easily interpreted by human, while frequency domain makes it easier to identify some periodic patterns, the frequency component. The two domains are equivalent in terms of describing the image.

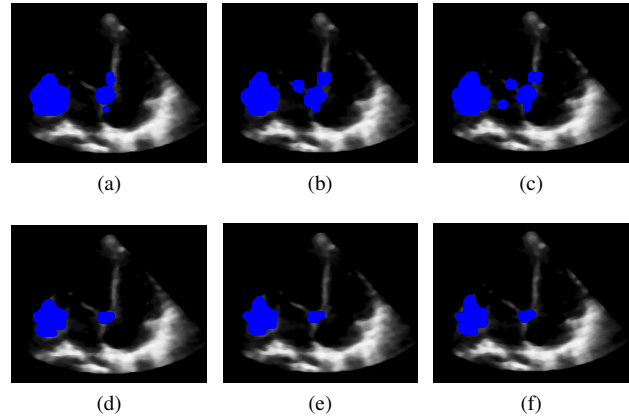


Fig. 3: Texture Descriptor

They just provides two different perspectives. One advantage of frequency domain is its insensitive to noise, which is usually sparse, and occupies the high frequency component. It is pretty easy task to eliminate such kind noise, and that makes it an effective tools dealing with signals corrupted by noise. In [5], Zhang and Lu draw a conclusion that Fourier descriptors outperform others. However, it also has some disadvantage. It is difficult to deal with image rotation, scaling and translation, due to the sensitivity of Fourier Transformation. To overcome the difficulty, a generic Fourier Descriptor (GFD) has been proposed by Zhang and Lu[8]. The GFD is acquired by applying a 2D Fourier transform on a polar-raster sampled image:

$$PF_2(\rho, \phi) = \sum_r \sum_i f(r, \theta) \exp \left[j2\pi \left(\frac{r}{R} \rho + \frac{2\pi i}{T} \phi \right) \right] \quad (4)$$

where $0 \leq r < R$ and $\theta_i = i(2\pi/T)$; $0 \leq i < T$, $0 \leq \phi < T$. R and T are the radical frequency resolution and angular frequency resolution respectively. The normalized coefficient are GFD. Zhang and Lu have shown that GFD outperforms other region-based descriptors such as 2D Fourier Descriptor and moments.

2.3 Classifier

The classifier we use here is Support Vector Machine (SVM). First proposed by Cortes and Vapnik[21] in 1995, SVM quickly became popular due to its simplicity and good performance. The common procedure is first training the classifier with known samples, and then the it is used for classification or regression. Mendizabal-Ruiz and Rivera [3] even use it to compute likelihoods. SVM is based on the concept of decision planes that separate data from different classes. The training of SVM involves minimization of the error function

$$\frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \quad (5)$$

subject to the constraints

$$\begin{aligned} y_i(w^T \phi(x_i) + b) &\geq 1 - \xi_i \quad \text{and} \\ \xi_i &\geq 0, \quad i = 1, \dots, N \end{aligned} \quad (6)$$

where C is the capacity constraint, w is the vector of coefficients, b is a constant, and ξ_i is the parameter to handle nonseparable data. The kernel function $\phi(x)$ is used to transform original data to feature space.

With kernel method, SVM gains the ability to deal with nonseparable cases by mapping sample features into high dimensional space, where the data may be linearly separated by a hyperplane. A lot of kernels have been developed in the literature, but linear, polynomial and Gaussian radial basis(RBF) belong to the most commonly used. In general, linear kernel should be tested first, because it is simple and efficient, and in many cases, linear kernel is good enough; then polynomial and then Gaussian radial basis function kernel. Each kernel is controlled by some hyper parameter, and a common technique for finding out the value of these hyper parameters is cross-validation and grid-search [4]. In our experiment we find linear kernel works for some case but fails for others, and polynomial kernel works just fine.

2.4 Post processing

After feature extraction and classification, we can make a through comparison of different descriptors. The result is presented in Section 3. To complete our task, some other steps are needed, but they are irrelevant to our evaluation, we would not discuss here.

3. Experiment

In this section, we present the details of our evaluation criterion and experiment result. The descriptors described above are evaluated on a set of cardiac ultrasound images of size 261×321 that come from hospital. Before start, we would like to discuss the evaluation criterion first.

3.1 Evaluation criterion

The evaluation criterion we use here is similar to one used in [7] that based on the number of correct and false recognition. To do this, we have manually labeled all the test images, denoted by \mathbf{A} . Then the classified result of test images, denoted by \mathbf{B} , are used to compare with the labeled images. The points labeled positive and classified as negative are said to be false negative, and those labeled negative but recognized positive are said to be false positive. The evaluation is presented with recall and 1-precision. Slightly different to [7], recall here describes the ratio of correct positive with respect to the total number of positive points, and is defined as:

$$recall = \frac{\#correct\ positive}{\#correct\ positive + \#false\ negative}$$

The number of correct positive can be defined by $\mathbf{A} \cap \mathbf{B}$, and the false negative as $\mathbf{A} \cap \mathbf{B}^c$, where \mathbf{B}^c is the complement of \mathbf{B} .

Unlike most vision applications where accuracy is the most important evaluation criterion, in our case, however, we are more concerned about 1-precision that measures the percentage of false positive. In the context of medical image, the penalty of a misclassification of abnormal to be normal is much severe than the opposite, because the former is easy to remedy through further inspection while the latter is difficult to detect and may cause great trouble. Thus, we care more about 1-precision that can be formulated as

$$1 - precision = \frac{\#false\ positive}{\#correct\ positive + \#false\ positive}$$

where correct positive is the same as recall and false positive is defined as $\mathbf{A}^c \cap \mathbf{B}$.

3.2 Experimental Result

As discussed in [4], we use cross-validation procedure to avoid overfitting problem and grid search algorithm to find the best parameter of SVM classifier. The dataset in this experiment comes from our cooperative hospital. There are 24 images for each series. Every image is manually labeled by experts and the labeling result is considered as the most authoritative reference and serves as a criterion to judge classification result by our trained classifier. The classifier we use here is a SVM-based library implemented by C.-J Lin, *LIBSVM* (see [4]). Lin provides several utility tools in his library, so that we can easily carry out the cross-validation and grid search procedure.

Our evaluation uses four parameters, recall, 1-precision, L and R, as their respective performance indicator. Parameter recall and 1-precision is explained in the evaluation criterion part and can be interpreted as something like precision and error rate, while L and R refer to the percentage that left and right root of mitral valve are recognized correctly, and should be considered as the overall correct rate. The result is listed in Table 1. All methods has a L of 1.0, showing that left root is relative easy to recognize, whereas only *Moment* has a R of 1.0 which means recognizing the right root is somewhat a challenging task. This is consistent with our intuition. We can see that GFD outperforms other descriptors when used alone, but it suffers from a little higher 1-precision. The combined descriptor performs better than individuals, both in recall and 1-precision.

4. Conclusion

In this paper, we present an evaluation of different descriptors on ultrasound images from hospital. Our goal is to find the most effective descriptors that can be used in the context of medical image processing. Since ultrasound images differ a lot from typical vision images. Experiments show that GFD outperforms other descriptors, followed by

Table 1: Experimental result

Description	recall	1 - precision	L	R
Context	0.933	0.088	1.0	0.833
Histogram	0.942	0.171	1.0	0.944
Histogram+Context	0.981	0.071	1.0	0.944
Moment	0.990	0.437	1.0	1.0
Texture	0.833	0.048	1.0	0.667
Texture+Context	0.944	0.073	1.0	0.833
GFD	0.967	0.108	1.0	0.944

texture. Combination of several simple descriptors tend to perform better than single descriptors.

References

- [1] K. Wu, H. Shu, and J. L. Dillenseger, "Region and boundary feature estimation on ultrasound images using moment invariants," *Computer methods and programs in biomedicine*, Vol. 113, pp. 446–455, Feb. 2014.
- [2] R. Rocha, J. Silva, and A. Campilho, "Automatic detection of the carotid lumen axis in B-mode ultrasound images," *Computer methods and programs in biomedicine*, Vol. 115, pp. 110–118, Jul. 2014.
- [3] E. G. Mendizabal-Ruiz, M. Rivera, and I. A. Kakadiaris, "Segmentation of the luminal border in intravascular ultrasound B-mode images using a probabilistic approach," *Medical image analysis*, Vol. 17, pp. 649–670, Aug. 2013.
- [4] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, paper 27, Apr. 2011.
- [5] D. Zhang, and G. Lu, "Review of shape representation and description techniques," *Pattern recognition*, Vol. 37, pp. 1–19, Jan. 2004.
- [6] Z. Tu, and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, pp. 1744–1757, Oct. 2010.
- [7] K. Mikolajczyk, and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, pp. 1615–1630, Oct. 2005.
- [8] D. Zhang, and G. Lu, "Shape-based image retrieval using generic Fourier descriptor," *Signal Processing: Image Communication*, Vol. 17, pp. 825–848, Nov. 2002.
- [9] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision*, Vol. 2, pp. 1150–1157, Sept. 1999.
- [10] Y. Ke, and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Computer Vision and Pattern Recognition, CVPR*, Vol. 2, pp. 506–513, Jul. 2004.
- [11] R. B. Yadav, N. K. Nishchal, A. K. Gupta, and V. K. Rastogi, "Retrieval and classification of shape-based objects using Fourier, generic Fourier, and wavelet-Fourier descriptors technique: A comparative study," *Optics and Lasers in engineering*, Vol. 45, pp. 695–708, Jun. 2007.
- [12] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp. 509–522, Apr. 2002.
- [13] A. Milkowski, Y. Li, D. Becker, and S. O. Ishrak, "Speckle reduction imaging," *Technical White Paper—General Electric Health Care (Ultrasound)*, Jul. 2009.
- [14] M. Yang, K. Kpalma, and J. Ronsin, "A survey of shape feature extraction techniques," *Pattern recognition*, pp. 43–90, 2008.
- [15] S. Loncaric, "A survey of shape analysis techniques," *Pattern recognition*, Vol. 31, pp. 983–1001, Aug. 1998.
- [16] G. Carneiro, and A. D. Jepson, "Phase-based local features," in *Computer Vision—ECCV 2002* pp. 282–296.
- [17] E. R. Davies, "On the noise suppression and image enhancement characteristics of the median, truncated median and mode filters," *Pattern Recognition Letters*, Vol. 7, pp. 87–97, Feb. 1988.
- [18] T. Lindeberg, "Feature detection with automatic scale selection," *International journal of computer vision*, Vol. 30, pp. 79–116, Nov. 1998.
- [19] C. G. Kiran, L. V. Prabhu, R. V. Abdu, and K. Rajeev, "Traffic sign detection and pattern recognition using support vector machine," in *ICAPR'09*, pp. 87–90, Feb. 2009.
- [20] J. V. Soares, J. J. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Transactions on Medical Imaging*, Vol. 25, pp. 1214–1222, Sept. 2006.
- [21] C. Cortes, and V. Vapnik, "Support-vector networks," *Machine learning*, Vol. 20, pp. 273–297, Sept. 1995.
- [22] T. C. Aysal, and K. E. Barner, "Rayleigh-maximum-likelihood filtering for speckle reduction of ultrasound images," *IEEE Transactions on Medical Imaging*, Vol. 26, pp. 712–727, May. 2007.
- [23] M. S. Nixon, and A. S. Aguado, *Feature Extraction & Image Processing for Computer Vision*, 3rd ed., London, UK: Academic Press, 2012.

NONCONVEX HALF-QUADRATIC IMAGE RECONSTRUCTION WITH WAVELET-DOMAIN SPARSITY CONSTRAINTS

Marc C. Robini and Yuemin Zhu

CREATIS (CNRS UMR 5220, INSERM U1044), INSA Lyon, 69621 Villeurbanne, France

ABSTRACT

Image reconstruction usually reduces to the minimization of an objective functional which combines data fidelity and edge-preserving regularization. We focus on the particularly difficult case where the data fidelity term is non-quadratic and the regularization is nonconvex and bounded. We propose to further stabilize the inverse problem by adding a wavelet-domain sparsity penalty to the spatial-domain regularization term, and we provide an efficient half-quadratic algorithm to solve the associated optimization task. Our approach generalizes previous work on half-quadratic image reconstruction and offers stronger convergence guarantees. Numerical experiments show that wavelet-domain sparsity regularization not only reduces optimization difficulty but also increases the quality of the reconstructions.

Index Terms— image reconstruction, edge preservation, inverse problems, nonconvex optimization, sparse regularization.

1. INTRODUCTION

We consider the problem of reconstructing a piecewise-smooth image with vector representation $\mathbf{x}^\sharp \in \mathbb{R}^N$ from its measurements

$$\mathbf{d} = \chi(\mathbf{D}\mathbf{x}^\sharp + \boldsymbol{\nu}), \quad (1)$$

where $\mathbf{D} \in \mathbb{R}^{M \times N}$ is the observation matrix, $\boldsymbol{\nu} \in \mathbb{R}^M$ is a noise vector whose components are realizations of independent zero-mean random variables, and $\chi : \mathbb{R}^M \rightarrow \mathbb{R}^M$ represents contamination by impulse noise. The set of solutions is usually defined as the set of minimizers of an objective functional $U : \mathbb{R}^N \rightarrow \mathbb{R}$ that combines a data fidelity term U_{fid} with a regularization term U_{reg} weighted by a parameter $\gamma > 0$:

$$U(\mathbf{x}) = U_{\text{fid}}(\mathbf{x}) + \gamma U_{\text{reg}}(\mathbf{x}). \quad (2)$$

The design of the data fidelity term is guided by the noise characteristics. The standard form is

$$U_{\text{fid}}(\mathbf{x}) = \sum_{m \in [1..M]} \theta_{\text{fid}}(|(\mathbf{D}\mathbf{x} - \mathbf{d})_m|), \quad (3)$$

where $(\mathbf{D}\mathbf{x} - \mathbf{d})_m$ denotes the m th component of the vector $\mathbf{D}\mathbf{x} - \mathbf{d}$ and $\theta_{\text{fid}} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is increasing and continuously differentiable. The regularization term favors images with smooth regions separated by edges; it is defined by

$$U_{\text{reg}}(\mathbf{x}) = \sum_{l \in [1..L]} \theta_{\text{reg}}(\|\mathbf{R}_l \mathbf{x}\|), \quad (4)$$

where $\theta_{\text{reg}} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is increasing and continuously differentiable, $\|\cdot\|$ is the ℓ_2 -norm, and $\mathbf{R}_l \in \mathbb{R}^{N \times \rho}$. Usually, $\{\mathbf{R}_l ; l \in [1..L]\}$ is a set of first-order difference operators ($\rho = 1$) or a discrete approximation to the gradient ($\rho = 2$). The function θ_{reg} is called a *potential function* (PF) (and so is θ_{fid} by extension); it must satisfy $\lim_{t \rightarrow +\infty} \theta_{\text{reg}}(t) < +\infty$ to preserve edges. By contrast with convex PFs which reduce smoothing at discontinuity locations [1], nonconvex PFs yield sharp edges [2] at the expense of high optimization difficulty.

We focus on the case where θ_{reg} is a bounded function such as, for example, the Tukey biweight function

$$\theta(t) = \begin{cases} 1 - (1 - t^2/(6\delta^2))^3 & \text{if } t \leq \sqrt{6}\delta, \\ 1 & \text{if } t > \sqrt{6}\delta, \end{cases} \quad (5)$$

where $\delta > 0$ is a scaling parameter that allows to adapt to the dynamic range of \mathbf{x}^\sharp . The boundedness of θ_{reg} strongly complicates the optimization task and increases the instability of the reconstruction process. We propose an improved model by introducing sparsity constraints in a multiresolution space. The objective functional is augmented by a penalty term $U_{\text{spa}}(\mathbf{x}) = \Psi(\mathbf{T}\mathbf{x})$, where $\mathbf{T} \in \mathbb{R}^{N \times N}$ is a non-redundant wavelet transform and Ψ is a sparsity penalty operating on the detail coefficients. We also propose an efficient half-quadratic algorithm to minimize the augmented cost functional. The associated convergence results extend those given in [1, 3, 4, 5, 6, 7]; they apply to a large class of objectives including convex and nonconvex functionals with a single minimizer, nonconvex functionals with isolated stationary points, and even some nonconvex functionals with non-isolated stationary points. Furthermore, we provide point-to-set and gradient convergence guarantees for the remaining case where a level set contains a bounded infinite set of stationary points.

This paper is organized as follows. Section 2 describes the wavelet-domain sparsity penalty, Section 3 presents the

optimization algorithm along with its convergence properties, and experimental results are given in Section 4.

2. WAVELET-DOMAIN SPARSITY REGULARIZATION

Let $\mathbf{T} \in \mathbb{R}^{N \times N}$ be a non-redundant, J -level wavelet transform. The vector $\mathbf{T}\mathbf{x}$ is the concatenation of the vectors in

$$\{\mathbf{T}_0\mathbf{x}\} \cup \left(\bigcup_{j \in [0..J]} \{\mathbf{T}_{j,h}\mathbf{x}, \mathbf{T}_{j,v}\mathbf{x}, \mathbf{T}_{j,d}\mathbf{x}\} \right), \quad (6)$$

where $\mathbf{T}_0\mathbf{x}$ is the approximation of \mathbf{x} at resolution 2^{-J} and $\mathbf{T}_{j,h}\mathbf{x}$, $\mathbf{T}_{j,v}\mathbf{x}$ and $\mathbf{T}_{j,d}\mathbf{x}$ are the detail images at level j representing the information lost when going from resolution 2^{j+1-J} to resolution 2^{j-J} . The operators $\mathbf{T}_{j,h}$, $\mathbf{T}_{j,v}$ and $\mathbf{T}_{j,d}$ give the high frequencies in the horizontal, vertical and diagonal directions, respectively, and each detail image at level j contains $4^{j-J}N$ wavelet coefficients (we implicitly assume that the original image $\mathbf{x}^\#$ is of size $N_1 \times N_2$ with N_1 and N_2 divisible by 2^J). Since piecewise-smoothness translates to sparse detail images, we propose to further regularize the reconstruction with the sparsity penalty

$$U_{\text{spa}}(\mathbf{x}) = \sum_{j \in [0..J-1]} \sum_{l \in [1..L_j]} \theta_{\text{spa}}(2^j |(\mathbf{T}_{j,x}\mathbf{x})_l|), \quad (7)$$

where the coefficient 2^j compensates wavelet-coefficient decay and $\mathbf{T}_{j,x} \in \mathbb{R}^{L_j \times N}$ ($L_j = 3 \cdot 4^{j-J}N$) is the vertical concatenation of $\mathbf{T}_{j,h}$, $\mathbf{T}_{j,v}$ and $\mathbf{T}_{j,d}$. The PF $\theta_{\text{spa}} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a smooth convex approximation to the identity on \mathbb{R}_+ — for example the Huber function

$$\theta(t) = \begin{cases} t^2/(2\delta^2) & \text{if } t \leq \delta, \\ 1 & \text{if } t > \delta, \end{cases} \quad (8)$$

where $\delta > 0$ must be small compared to the dynamic range of $2^j \mathbf{T}_{j,x}\mathbf{x}$ for every $j \in [0..J-1]$.

The new solutions to the reconstruction problem are the minimizers of the augmented objective functional

$$V(\mathbf{x}) = U_{\text{fid}}(\mathbf{x}) + \gamma U_{\text{reg}}(\mathbf{x}) + \tilde{\gamma} U_{\text{spa}}(\mathbf{x}), \quad (9)$$

where $\tilde{\gamma} > 0$ controls the strength of the sparsity penalty. The advantages of this approach are the following:

1. The augmented functional V has the same form as the original functional U ; so we can use the same strategy for minimizing either U or V .
2. There is only one additional free parameter (namely $\tilde{\gamma}$).
3. There is no increase in optimization difficulty since the sparsity penalty is convex.
4. The sparsity penalty excludes spurious solutions \mathbf{x}^* such that $\|\mathbf{D}\mathbf{x}^* - \mathbf{d}\| \approx 0$ and $U_{\text{reg}}(\mathbf{x}^*) \approx \sup U_{\text{reg}}$ (recall that U_{reg} is bounded), which are not piecewise-smooth.

3. HALF-QUADRATIC OPTIMIZATION

The augmented functional V has the general form

$$V(\mathbf{x}) = \sum_{k=[1..K]} \gamma_k \theta_k(\|\mathbf{A}_k\mathbf{x} - \mathbf{a}_k\|), \quad (10)$$

where $\gamma_k > 0$, $\mathbf{A}_k \in \mathbb{R}^{N_k \times N}$ and $\mathbf{a}_k \in \mathbb{R}^{N_k}$. We assume throughout that every PF $\theta_k : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ satisfies the following conditions:

- θ_k is increasing and nonconstant,
- θ_k is C^1 on $(0, +\infty)$ and continuous at zero,
- $t^{-1}\theta'_k(t)$ is decreasing and bounded on $(0, +\infty)$.

(Note that θ_k can be bounded and even eventually constant.) It follows that θ'_k is right-differentiable at zero and $\theta''_k(0) = \lim_{t \rightarrow 0^+} t^{-1}\theta'_k(t)$. For every k , we let $\omega_k : \mathbb{R}^N \rightarrow \mathbb{R}$ be the functional defined by

$$\omega_k(\mathbf{x}) = \theta_k^\dagger(\|\mathbf{A}_k\mathbf{x} - \mathbf{a}_k\|) \quad (11)$$

$$\text{with } \theta_k^\dagger(t) = \begin{cases} t^{-1}\theta'_k(t) & \text{if } t > 0, \\ \theta''_k(0) & \text{if } t = 0. \end{cases} \quad (12)$$

The half-quadratic optimization algorithm is the following.

Algorithm 1 Given a starting point $\mathbf{x}^{(0)} \in \mathbb{R}^N$, generate the sequence $(\mathbf{x}^{(p)})_{p \in \mathbb{N}}$ by using the recurrence relation

$$\mathbf{x}^{(p+1)} = \arg \min_{\mathbf{y} \in \mathbb{R}^N} Q(\mathbf{x}^{(p)}, \mathbf{y}) \quad (13)$$

$$\text{with } Q(\mathbf{x}, \mathbf{y}) = \sum_{k \in [1..K]} \omega_k(\mathbf{x}) \|\mathbf{A}_k\mathbf{y} - \mathbf{a}_k\|^2. \quad (14)$$

Each iteration consists in computing the weights $\omega_k(\mathbf{x}^{(p)})$ and then minimizing the quadratic functional $Q(\mathbf{x}^{(p)}, \cdot)$. So algorithm 1 is well-defined if and only if $Q(\mathbf{x}, \cdot)$ is positive definite for every \mathbf{x} , or equivalently,

$$\bigcup_{k \in \mathcal{J}_+} \ker(\mathbf{A}_k) = \{\mathbf{0}\}, \quad (15)$$

where \mathcal{J}_+ is the set of indices k such that θ_k is strictly increasing. For example, if the spatial-domain regularization term U_{reg} in (9) is defined by an eventually constant PF θ_{reg} , then condition (15) writes

$$\ker(\mathbf{D}) \cap \ker(\mathbf{T}_{\setminus 0}) = \{\mathbf{0}\}, \quad (16)$$

where $\mathbf{T}_{\setminus 0}$ is the vertical concatenation of the all the detail operators.

Algorithm 1 has different interpretations: alternating minimization [1, 4], majorization-minimization [6], quasi-Newton optimization [6, 8], and fixed-point iteration [7]. In the following subsections, we give global convergence results that generalize those in the above references. The proofs are omitted because of space limitation and can be found in [9].

3.1. Convergence in norm

Let $\mathcal{G} \subseteq \mathbb{R}^N$ be nonempty. A point $\mathbf{x} \in \mathcal{G}$ is said to be *isolated* (in \mathcal{G}) if \mathcal{G} is a singleton or if there is an $\alpha > 0$ such that $\|\mathbf{y} - \mathbf{x}\| \geq \alpha$ for all $\mathbf{y} \in \mathcal{G} \setminus \{\mathbf{x}\}$. The set \mathcal{G} is called *discrete* if all its points are isolated. Let \mathcal{S} denote the set of stationary points of V , that is,

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^N : \nabla V(\mathbf{x}) = \mathbf{0}\}. \quad (17)$$

We call \mathcal{S} *level-discrete* if, for every $h \in \mathbb{R}$, the set $\{\mathbf{x} \in \mathcal{S} : V(\mathbf{x}) = h\}$ (i.e., the intersection of \mathcal{S} with the h -level set of V) is either discrete or empty. We can construct examples of objective functionals satisfying our conditions (including (15)) and such that \mathcal{S} is level-discrete but not discrete.

Theorem 1 below gives conditions for convergence to a stationary point whether or not V is convex.

Theorem 1 *Let $(\mathbf{x}^{(p)})_p$ be a sequence generated by Algorithm 1 under condition (15), and assume that the set*

$$\mathcal{E}_0 = \{\mathbf{x} \in \mathbb{R}^N : V(\mathbf{x}) \leq V(\mathbf{x}^{(0)})\} \quad (18)$$

is bounded. If \mathcal{S} is level-discrete, then $(\mathbf{x}^{(p)})_p$ converges to a point in \mathcal{S} . In particular, if V has a unique stationary point \mathbf{x}^ , then \mathbf{x}^* is the unique global minimizer of V and $(\mathbf{x}^{(p)})_p$ converges to \mathbf{x}^* .*

If V is coercive (i.e., $V(\mathbf{x}) \rightarrow +\infty$ as $\|\mathbf{x}\| \rightarrow +\infty$), then \mathcal{E}_0 is bounded whatever the starting point. Furthermore, we can show that V is coercive if and only if $\bigcup_{k \in \mathcal{J}_\infty} \ker(\mathbf{A}_k) = \{\mathbf{0}\}$, where \mathcal{J}_∞ is the set of indices k such that $\theta_k(t) \rightarrow +\infty$ as $t \rightarrow +\infty$. So coercivity also implies (15), which leads to the following corollary.

Corollary 2 *If V is coercive, then any sequence generated by Algorithm 1 converges to a point in \mathcal{S} .*

Theorem 1 can be further specialized to the case where V is strictly convex (which is possible even if some of the PFs θ_k are eventually constant), for then V is coercive and has a unique stationary point.

Corollary 3 *If V is strictly convex, then any sequence generated by Algorithm 1 converges to the unique global minimizer of V .*

3.2. Point-to-set and gradient convergence

The following theorem shows that Algorithm 1 also behaves well when \mathcal{S} is not level-discrete.

Theorem 4 *Let $(\mathbf{x}^{(p)})_p$ be a sequence generated by Algorithm 1 under condition (15), and suppose \mathcal{E}_0 is bounded.*

(i) *If $(\mathbf{x}^{(p)})_p$ does not converge to a point in \mathcal{S} , then*

$$\lim_{p \rightarrow \infty} \inf_{\mathbf{y} \in \mathcal{S}^\circ} \|\mathbf{y} - \mathbf{x}^{(p)}\| = 0, \quad (19)$$

where \mathcal{S}° denotes the interior of \mathcal{S} .

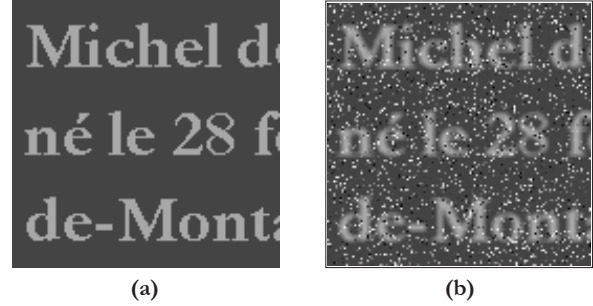


Fig. 1. “Text” image: (a) original; (b) degraded observation (3×3 uniform blur + 30 dB white Gaussian noise + 15% random-valued impulse noise).

$$(ii) \lim_{p \rightarrow \infty} \|\nabla V(\mathbf{x}^{(p)})\| = 0.$$

In other words, Algorithm 1 converges either to a stationary point in the usual sense or to the boundary of \mathcal{S} in terms of point-to-set distance, and the gradient of the objective always goes to zero.

If V is convex and coercive, then \mathcal{S} is the set of global minimizers of V (and is closed and convex). Hence the following corollary.

Corollary 5 *If V is convex and coercive, then any sequence generated by Algorithm 1 converges to the boundary of the set of global minimizers of V .*

4. EXPERIMENTS

We consider the reconstruction of the 128×128 image \mathbf{x}^\sharp shown in Fig. 1(a) from the data displayed in Fig. 1(b). The data were generated by blurring with a 3×3 uniform mask, then adding white Gaussian noise at 30 dB SNR, and finally adding random-valued impulse noise with a 15% corruption rate. The original pixel intensity is either 50 (background) or 150 (text regions), and the data has intensity values ranging from 0 to 255. We assess reconstruction quality via the improvement in SNR (ISNR) (see, e.g., [7]).

The estimates of \mathbf{x}^\sharp are obtained by minimizing either U (2) or V (9) using Algorithm 1. We use the same PF for data fidelity and wavelet-domain sparsity regularization: θ_{fid} and θ_{spa} are set to be the Huber function (8) with $\delta = 0.1$ (so considering the dynamic range of \mathbf{x}^\sharp , both PFs are close approximation to the identity on \mathbb{R}_+). The spatial-domain regularization PF θ_{reg} is the Tukey function (5) with $\delta = 20\sqrt{6}$, and $\{\mathbf{R}_l; l \in [1 \dots L]\}$ is a discrete approximation to the gradient; therefore, $\theta_{\text{reg}}(\|\mathbf{R}_l \mathbf{x}\|) = 1$ at all locations where the gradient magnitude is greater than 20. As regards wavelet-domain sparsity regularization, we use a biorthogonal spline wavelet transform with one resolution level (i.e., $J = 1$).

If we use spatial regularization only, condition (15) becomes $\ker(\mathbf{D}) = \{\mathbf{0}\}$. But \mathbf{D} is a uniform-blur operator and so Algorithm 1 is not well-defined for U . We can get around

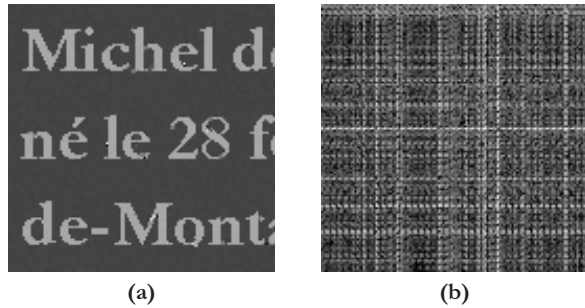


Fig. 2. Reconstruction of the “text” image without sparsity regularization: (a) best result obtained by varying γ (ISNR = 13.46 dB); (b) deep spurious minimizer displayed in absolute value using a square-root gray-scale (ISNR = -65 dB).

this problem by letting the PF θ_{reg} vary with p to reach a basin of attraction before using the Tukey function: we use a sequence of the form $\varepsilon_p \theta_T + (1 - \varepsilon_p) \theta_H$, where θ_T and θ_H are the Tukey and Huber functions and ε_p increases linearly from 0 to 1 in 50 iterations. Figure 2(a) displays the result obtained by minimizing U with the best setting of the regularization strength γ . The quality of this particular estimate—call it \mathbf{x}^* —is satisfactory, but the reconstruction process is unstable because U has deep spurious minimizers. For example, the image \mathbf{y}^* shown in Fig. 2(b) is a minimizer of U for the same value of γ as that used to compute \mathbf{x}^* . Although \mathbf{y}^* is much farther away from $\mathbf{x}^\#$ than is \mathbf{x}^* , it is much deeper: $U(\mathbf{y}^*) = 5.69 \cdot 10^4$ whereas $U(\mathbf{x}^*) = 2.01 \cdot 10^5$.

The addition of the wavelet-domain sparsity penalty U_{spa} has two immediate advantages. First, Algorithm 1 is well-defined for V (indeed, condition (16) holds since the images in $\ker(\mathbf{D})$ have strong high-frequency content while $\ker(\mathbf{T}_{\setminus 0})$ contains smooth interpolations of low resolution images). Second, the reconstruction process is stabilized because U_{spa} favors sparse edge-maps and strongly penalizes spurious minimizers of U . Figure 3(a) displays the ISNR as a function of the sparsity regularization strength $\tilde{\gamma}$ when γ is fixed to the value that led to the estimate in Fig. 2(a). The best reconstruction obtained is shown in Fig. 3(b); the corresponding ISNR is about 6 dB higher than that achieved without sparsity regularization. For large values of $\tilde{\gamma}$ (greater than 2 in the present case) the ISNR stabilizes at about 8 dB and the solution is a low-pass filtered version of $\mathbf{x}^\#$. Figure 4 compares the ISNR versus γ curves obtained with and without sparsity regularization ($\tilde{\gamma}$ is fixed to the value that led to the estimate in Fig. 3(b)). We observe that sparsity regularization systematically increases the range of values of γ for which the ISNR is above a given threshold—in other word, the reconstruction is more stable.

5. REFERENCES

[1] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, “Deterministic edge-preserving regularization in computed imaging,” *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 298–311, 1997.

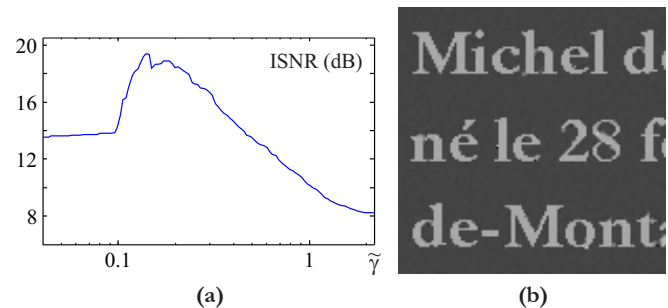


Fig. 3. Reconstruction of the “text” image with wavelet-domain sparsity regularization: (a) ISNR as a function of $\tilde{\gamma}$ when γ is fixed; (b) best result obtained by varying $\tilde{\gamma}$ (ISNR = 19.42 dB).

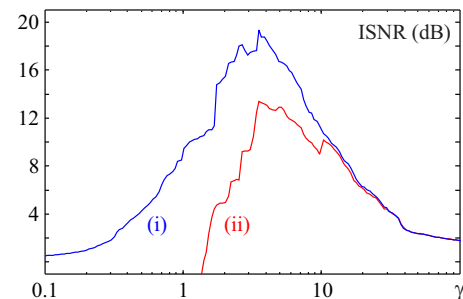


Fig. 4. ISNR as a function of γ : (i) with wavelet-domain sparsity regularization and a fixed $\tilde{\gamma}$; (ii) without sparsity regularization.

- [2] M. Nikolova, “Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares,” *Multiscale Model. Simul.*, vol. 4, no. 3, pp. 960–991, 2005.
- [3] A. H. Delaney and Y. Bresler, “Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography,” *IEEE Trans. Image Process.*, vol. 7, no. 2, pp. 204–221, 1998.
- [4] J. Idier, “Convex half-quadratic criteria and interacting auxiliary variables for image restoration,” *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1001–1009, 2001.
- [5] M. Nikolova and M. Ng, “Analysis of half-quadratic minimization methods for signal and image recovery,” *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.
- [6] M. Allain, J. Idier, and Y. Goussard, “On global and local convergence of half-quadratic algorithms,” *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1130–1142, 2006.
- [7] M. C. Robini, Y. Zhu, and J. Luo, “Edge-preserving reconstruction with contour-line smoothing and non-quadratic data-fidelity,” *Inverse Probl. Imaging*, vol. 7, no. 4, pp. 1331–1366, 2013.
- [8] M. Nikolova and R. Chan, “The equivalence of half-quadratic minimization and the gradient linearization iteration,” *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1623–1627, 2007.
- [9] M. C. Robini, “Generic half-quadratic optimization for signal reconstruction,” available on request by e-mail to marc.robini@creatis.insa-lyon.fr, 2014.

A 3-D auto-stereoscopic integral images generating tools

Shafik Salih, Amar Aggoun, Maysam Abbod

Abstract - in this paper, application software devoted to simulate a 3-D integral imaging system (3-D camera) and produce static and animated 3-D autostereoscopic good quality integral images that are viewable without the need for special glasses is introduced. The resulting images are capable to provide the viewer with 3-D effect and parallax. A Human Machine Interface (HMI) is provided with the software to allow user to tune and set the parameter values of the required imaging systems and the required integral image characteristics. The software receives the instructions and the parameter values from the HMI, and imports the computer-generated 2-D scenes that are intended to be rendered as 3-D integral images. The integral imaging process that is implemented with the simulated camera is based on particular algorithms introduced for this purpose (e.g Displacement of Camera Target *DCTarget* and the integral imaging method of dividing the camera view volume *DIVGL*).

Keywords: Application software, microlenses array.

1 Introduction

IMAGES with 3-D effect are more realistic and closer to the real world images as they create an influence on the human eyes similar to that created by real objects. In general, 3-D images and videos are preferred over the traditional 2-D images. There is growing evidence that 3-D imaging techniques will have the potential to establish a future mass-market in the fields of entertainment (television, video game) and communications (desktop video conferencing) [1].

Adding a third dimension to the traditional 2-D images and videos was the subject of many different approaches and inventions. 3-D images and videos of several types require the viewers to wear special glasses to be able to view the 3-D effect of these images. The need for such devices can form a practical obstacle in the cases when the viewer is not prepared to watch 3-D images. For example, if the 3-D images are meant to be used for outdoor advertisements. In order to implement free viewing 3-D display, different methods were introduced aiming to produce 3-D autostereoscopic images.

Despite that the computer generation of integral images was the subject of several studies, but the application software

that was created to build 3-D static and animated integral images and discussed in this study has the features that make it flexible, fast, accurate, and useful for general usage. Olsson and Xu [7] created simulation environment to allow for a simple definition of complex scenes and from those, integral images are synthesized. The application software and the Graphical User Interface (*GUI*) explained in this paper are simpler and more flexible as a wider range of 2-D static and animated scenes can be converted immediately to images and video with 3-D perception. Milnthorpe designed a software model to simulate Davies and McCormick system to render static and dynamic images in integral format [5] [10]. The algorithms used in our software have advantages over that method. With our software, unlimited integral imaging systems can be simulated providing the user with the ability to tune the imaging system parameters such as the microlenses pitch, microlenses focal length, the number of microlenses, and the type of microlenses used as a virtual microlenses array in the capture stage. Better image quality can be produced with a faster, more accurate, and more professional rendering process. The outputted images are integral images with either horizontal or vertical and horizontal parallax.

The application software that is supplied with a user-friendly interface and capable to produce autostereoscopic integral images is the subject of this study in which the stages needed for producing the integral images from the start to the end are explained. The software tool that is utilized to implement the methods of Displacement of Camera Target (*DCTarget*), and Dividing the Image Volume (*DIVGL*) is the plug-in tool or the application software that is intended to simulate 3-D cameras and produce 3-D integral images. Hardware devices or physical tools such as a PC and a microlenses array are employed to generate the required integral images and display the resulting images. The *GUI* allows the user to build the scene from the computer-generated models and select the measurements, the parameters, and the features of the scene. In addition, the *GUI* should allow the user to select the features and the parameters of the virtual camera and display devices, and then obligate the application to generate the integral images with the required features and parameters based on the *DCTarget* or *DIVGL* method.

11-11-2014

This work was supported by Brunel University, London. S. Salih is with the School of Engineering and Design, Brunel University, Uxbridge UB8 3PH, U.K.

2 The integral images production system structure

Figure 1 depicts the stages of a simple integral images production system, the interacted blocks and the interactions between these blocks. In order to produce the integral image of a scene, the scene components including the models and textures should be generated and saved in the PC memory beforehand. The scene components should be imported to the application environment providing the format of the files that hold the scene geometric data or the formats of the texture image files are supported by the implemented plug-in tool. The application environment is a Visual C++ supported by specialist libraries including Irrlicht library. In the same environment, a user-friendly *GUI* is used to control the scene and the integral image rendering. Through the *GUI*, the model and texture files are selected to build the static or animated scenes that are going to be converted to 3-D autostereoscopic integral images.

The output of the rendering process is a saved image file for a static scene, or a collection of image files for an animated scene so that each saved image file is the integral image of a single frame of the animated scene. The resulting image files can be displayed on a normal PC screen or a special display screen and viewed through a suitable microlenses array. The produced integral images would be viewed as a scene provided with a 3-D autostereoscopic effect. In addition, a horizontal parallax would appear when using the cylindrical microlenses array as the rendering mode, whereas, a horizontal and vertical parallax would appear when the spherical microlenses array mode is selected.

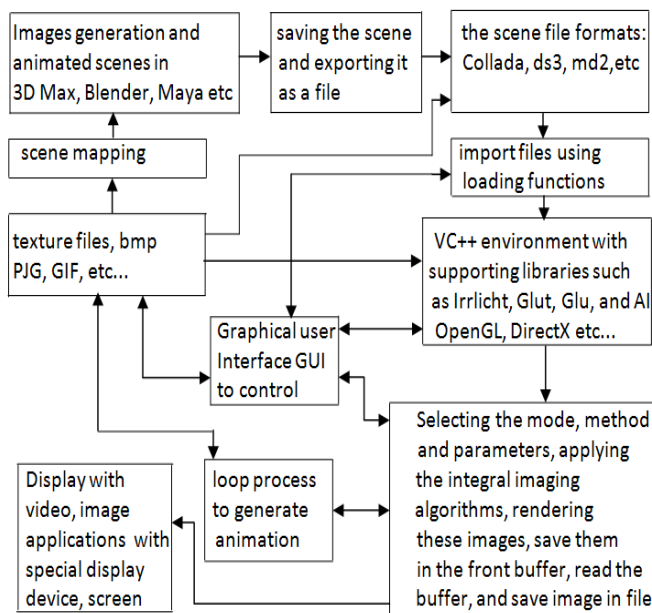


Fig. 1. The integral images production stages.

3 The application components and integral images production stages

3.1 Three-Dimension images generation

The integral images that are produced by the application are identical to the integral images produced by a simulated virtual 3-D camera. Therefore, the images that can be converted to 3-D autostereoscopic integral images using the application software are computer-generated images. The whole scene can be generated or its components beforehand. User should be able to build the scene and add new features to the view through the *GUI*. The scene components can be generated using applications such as 3D Max, Blender and Maya. The application should be able to import any file with a format supported by the application. The image files that hold the information needed for texture mapping should be available for rendering the models and the scene components. Specific texture file formats are supported.

3.2 Import 3-D images

When the model file or the model frames that are forming a scene are stored in files with certain file formats, and the texture files are saved in files with certain image file formats, these files should be loaded to the environment of the application software (e.g. VC++). Based on the selected file, the correct loading function is called. Irrlicht library provides such functions that are called to load files with the supported file formats including the texture image files.

3.3 Integral images production algorithm

Once the scene files including the geometric files and the texture files are loaded, an integral image production algorithm is applied on the scene to produce and display the required autostereoscopic images. The algorithms called *DCTarget* and *DIVGL* are used to generate the required 3-D autostereoscopic integral images.

3.4 Integral images display devices

Once the integral images are rendered and saved in files, these images can be displayed on a screen (e.g. PC screen) mounted by the correct microlenses array to view the scene with the required 3-D effect and parallax. The higher screen resolution we use the better image quality we get. The microlenses array can be cylindrical if the selected rendering mode is cylindrical or spherical when the rendering mode is spherical. Several factors can affect the quality of the displayed integral image such as the microlens angle of view.

A special display screen supplied with the required cylindrical or spherical microlenses array can be introduced. The images should be adjusted and displayed on such a screen so that the micro images of the integral images match the microlenses of the screen microlenses array.

3.5 Graphical User Interface

In order to optimize the characteristics of the resulting integral images, the Graphical User Interface (*GUI*) is needed to allow the user to tune the different parameters of the simulated imaging system. The parameters toolset should appear to allow the user to set the parameters. Figure 2 shows an example of a simple toolset and a default 2-D model. The *GUI* should allow the user to modify the scene and build a complete scene from a number of model files, textures, and light sources. From the menu, the required scene files and textures can be picked up to add objects and elements to the scene, and then, the parameters of each object would be modifiable. The other elements such as the light sources and the camera position and target should be controllable.



Fig. 2. A simple GUI based on Irrlicht engine.

4 Flow chart and sequential process

Figure 3 shows a diagram of the steps implemented by the plug-in tool to create the main control tools of the interface shown in Figure 2 and based on Irrlicht engine. When running the program, an event receiver is created, OpenGL driver is called, and a device is created. Using functions defined in the created device class, XML reader, scene manager, and video driver are created, also, models saved in archive files can be added to the scene. From the created device, an environment is defined. With the functions provided in the environment class, a root element from which a dialog windows can be created as well as toolboxes with their labels. The environment includes what the user can view of the control tools. The user interacts with the *GUI* using these tools. The functions provided with the environment class are called to create a menu and its submenus, tab controls, tabs, the combo boxes called Edit Boxes, toolbars, buttons, static texts, scrollbars, messages boxes and static images such as Brunel logo.

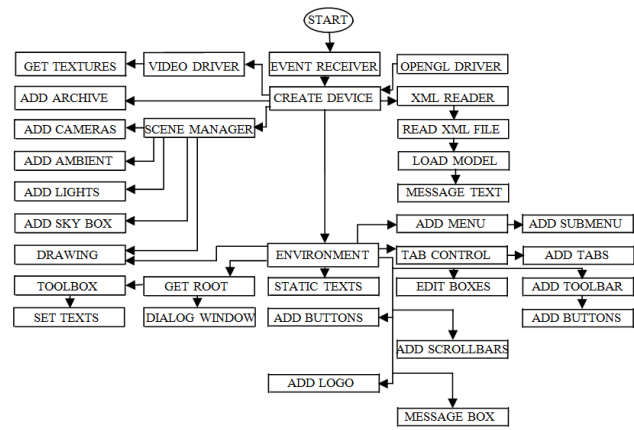


Fig. 3. A diagram of the steps implemented to create interface control tools.

Figure 4, depicts a flowchart of the process implemented by the plug-in tool to produce static and animated integral images based on the *DCTarget* algorithm. An event receiver is created to receive the user actions and convert them to convenient values for using them in the program. The model and texture files included in the XML files are loaded initially and displayed on the screen. The environment and input control are created to provide the interface needed to enter the scene and devices parameters manually.

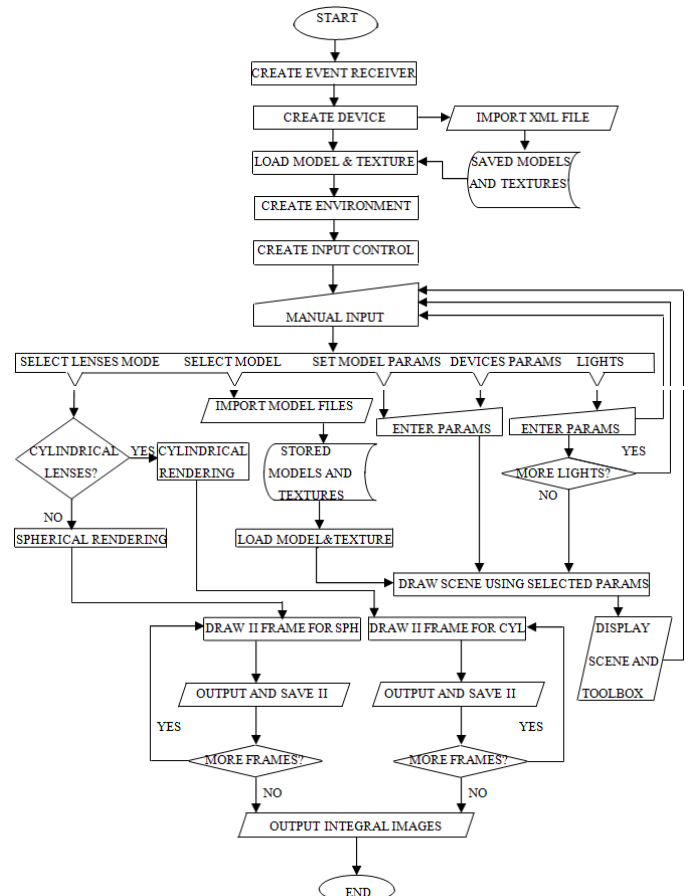


Fig. 4. A flowchart of the integral imaging process.

The microlenses array mode used to capture and display the scene can be selected. The resulting 3-D autostereoscopic integral image for each static 2-D image or for each frame from an animated scene is saved in a specific location in the memory. When all the frames are rendered, the process is ended, and a collection of integral images is saved. In order to build the scene, from the menu bar, the required models and textures saved in files are selected, imported to the scene environment and loaded; in addition, other scene elements are added. The loaded scene should be displayed instantly on the screen and multiple models should be able to be added to the scene at the same time, then the scene that is displayed on the screen is ready to be modified manually through the *GUI*. The new model parameters affect the displayed scene instantly and their values are used later in rendering the integral image.

5 Examples and results

Figure 5 shows an example when the interface window and a 2-D scene that is going to be modified, added another model and texture, and then rendered as a 3-D autostereoscopic integral image, the cylindrical microlenses mode was selected.

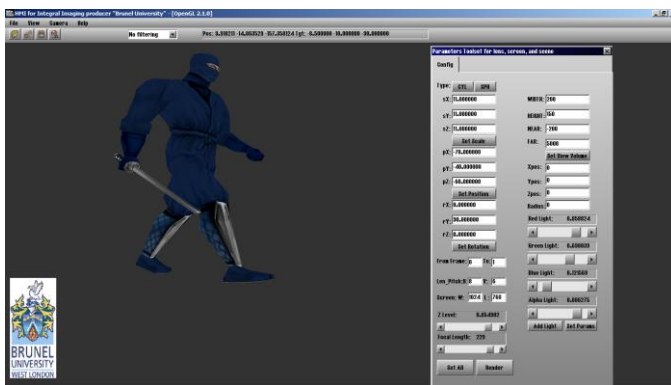


Fig. 5. A 2-D scene to be modified and rendered in the cylindrical mode.

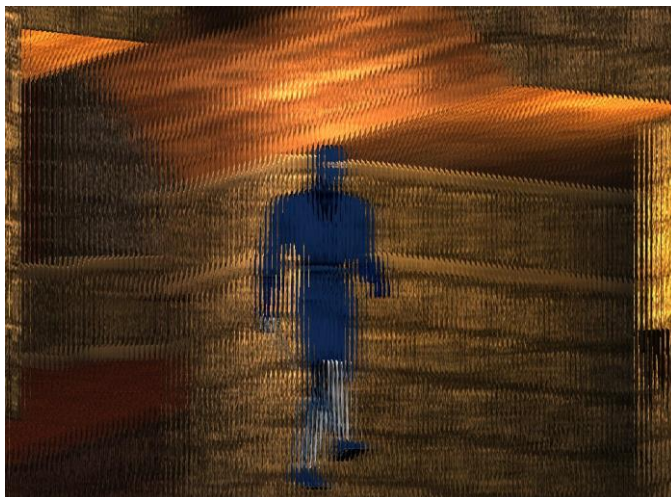


Fig. 6. An integral image resulting when using a cylindrical microlenses array.

The scene is instantly displayed in a 2-D mode to allow the user to test and tune the devices parameters and modify the scene if that is needed. Figure 6 shows the resulting integral image. Figure 7 shows the model when it is viewed through a primitive cylindrical microlenses array with a low quality.



Fig. 7. An integrated image as it is seen using a primitive microlenses array.

6 Conclusions

With the explained simple plug-in tool, 3-D cameras can be simulated and the parameters of the camera and the scene features can be selected and modified with the provided *HMI*. Static and animated integral images can be rendered using the plug-in tool. However, more developments are needed to enhance the abilities of the application and as a result the quality of the rendered integral images.

7 Acknowledgments

The author would like to thank Dr Emmanuel Tseklevs for his help and support, and the School of Engineering and Design at Brunel University for the financial support.

8 References and bibliography

- [1] A. Aggoun, "Pre-processing of integral images for 3-D displays," *Journal of Display Technology*, vol. 2, no. 4, Dec 2006.
- [2] Donald Hearn, M. Pauline Baker, "Computer Graphics with OpenGL", 3rd edition, *Published by Pearson Prentice Hall*, USA 2010.
- [3] <http://www.opengl.org/>, the formal OpenGL website. *Access date: 2012*.
- [4] C. V. Berkel, J. A. Clarke, Philips Research Laboratories, UK, "Characterisation and optimisation of 3D-LCD module design", *SPIE Digital Library*, vol. 3012, 2006.
- [5] G. E. Milnthorpe, "Computer generation of integral images using interpolative shading techniques", PhD thesis, School of Engineering and Manufacture, De Montfort University, Leicester, November 2003.
- [6] J. Kassebaum, N. Bulusu, W. Feng, "Smart camera network Localization using a 3D target", Portlan State University, Sep 2009.
- [7] E. B. Javidi and F. Okano, *Three-Dimensional Television, video, and Display Technologies*, Springer, 2002.
- [8] Aggoun, "3D Holoscopic Imaging Technology for Real-Time Volume Processing and Display", School of Engineering and Design, Brunel University, Uxbridge, UB8 3PH, (UK), 2010.
- [9] A. Aggoun, "3D Visual information engineering (3D VIE)", School of Engineering and Design, Brunel University, UK, 2011.
- [10] G. Milnthorpe, M. McCormick, N. Davies, "Computer modelling of lens arrays for Integral Image rendering," Imaging Technologies Group, SER Centre, De Montfort University, Leicester LE1 9BH UK, May, 2006.

SEM Surface Reconstruction using optical flow and Stereo Vision

J. C. Henao¹, J. Meunier², J. B. Gómez-Mendoza³ and J. C. Riaño-Rojas⁴

¹Department of Physics and Chemistry, Universidad Nacional de Colombia, Manizales, Caldas, Colombia

²Department of Computer Science and Operations Research,
Université de Montréal, Montréal, Québec, Canada

³Department of Electric, Electronic and Computing Engineering,
Universidad Nacional de Colombia, Manizales, Caldas, Colombia

⁴Department of Mathematics and Statistics,
Universidad Nacional de Colombia, Manizales, Caldas, Colombia

Abstract—*Scanning Electron Microscopy (SEM) is a tool used in a broad range of scientific and engineering applications. For instance, in materials science it is used to characterize the surface roughness and wear of materials. Despite the capability of SEM to obtain three-dimensional-like images, further manipulation of the images is necessary to obtain real three-dimensional information. In this paper, we present a fully automated method based on the differential technique for estimating Optical Flow for 3D reconstruction from a stereo pair. Stereo pairs are acquired by tilting a specimen a few degrees depending of the characteristics of the surface. With this methodology we obtained a good 3D approximation for different kinds of surfaces revealing valuable three-dimensional information.*

Keywords: SEM, Stereo Vision, Optical Flow, 3D Reconstruction.

1. Introduction

Scanning Electron Microscopy is a useful tool to obtain measures, characteristics and surface information on a nanometer (nm) to micrometer (μm) scale [3]. The possibility of obtaining three-dimensional-like images of the surfaces using SEM is very appreciated. This feature makes SEM very popular in a wide variety of media, from scientific journals, to popular magazines, and movies, etc.

In general, a SEM apparatus is composed by an electron column, a sample chamber, detectors of different kinds, and a visualization system [3], [4]. In SEM, a fine beam of electrons with high energies [1] is focused on a specimen, sweeping along a pattern of parallel lines. Different kinds of signals are generated as a result of the impact of the incident electrons, which are collected to form an image or to analyze the sample surface [2]. These are mainly secondary electrons, backscattered electrons and X-rays [3], [4].

SEM gives information about surface topography, crystalline structure, chemical composition and electrical behavior [2].

Although some three-dimensional information appears in a single SEM image [3], [5], [6], it is necessary to implement more sophisticated techniques to recover the 3D roughness of a surface. An interesting way to obtain 3D information is to use techniques that rely on stereo-vision. Two images (stereo pair) are taken with a small tilt angle difference over approximately the same area with a SEM [5].

In this paper, a technique for approximating surface height using Optical Flow based stereo reconstruction is presented. Corresponding points in each stereo pair are matched using the SURF algorithm [10], and then a geometric transformation is performed in order to correct pair misalignments. Optical Flow between the images is estimated using the differential approach. Results of applying the method to specimens of different materials are presented.

2. Fundamentals

2.1 Optical Flow

Optical Flow is the name given to the visible two-dimensional image motion which result from the projection of the three-dimensional motion of objects relative to a visual sensor in the image plane. Optical Flow can also be viewed as the pattern of apparent movement of objects, surfaces and edges, caused by the relative movement of the observer looking at the scene (or vice versa) [7].

Optical Flow estimation can be carried out using a differential approach, in which case the process typically involve three steps [9]

- Image smoothing with low-pass/band-pass filters, aiming to reduce the effect of image noise in derivative approximation.
- Spatial and time derivative approximations.
- Integration of the calculated derivatives in order to produce a 2D flow field.

In this work, velocity is computed from spatiotemporal derivatives of image intensity. A requirement in the application of differential techniques for Optical Flow estimation is

that $I(x, y)$ must be differentiable [9], condition that is met thanks to the use of a Gaussian kernel for image smoothing.

For a $2D + t$ case, a pixel placed at (x, y, t) with an intensity $I(x, y, t)$, moving $\Delta x, \Delta y$ in Δt between two frames of a scene, and under the assumption that the intensity is preserved in the image pair, the brightness constancy constraint is given by [7], [8], [9]

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \quad (1)$$

Assuming that the movement is small, and given the brightness constancy constraint in $I(x, y, t)$, local derivatives can be approximated using Taylor series to obtain

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\partial I}{\partial x} \Delta x + \frac{\partial I}{\partial y} \Delta y + \frac{\partial I}{\partial t} \Delta t \quad (2)$$

From equations 1 and 2 we have

$$\frac{\partial I}{\partial x} \Delta x + \frac{\partial I}{\partial y} \Delta y + \frac{\partial I}{\partial t} \Delta t = 0$$

or

$$\frac{\partial I}{\partial x} \frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y} \frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} = 0$$

During SEM stereo pair acquisition, we impose the constraint that the tilt of the specimen is around Y axis, so there is no movement in such direction ($\partial I / \partial y = 0$). Next, defining $\partial I / \partial x = I_x$, $\partial I / \partial t = I_t$ and $\Delta x / \Delta t = v_x$ leads to

$$I_x v_x = -I_t$$

or

$$v_x = -\frac{I_t}{I_x} \quad (3)$$

If we set $\Delta t = 1$, then $v_x = \Delta x$, which corresponds to the horizontal motion that will be used later to compute the disparities between two stereo images in order to infer the 3D surface relief.

3. Methodology

3.1 SEM Images Acquisition

A SEM stereoscopic pair is acquired by tilting the specimen a few degrees and capturing approximately the same region of interest [5]. For flatter surfaces the rotation has to be larger. It is fundamental to restrict the rotation to happen only around one axis to produce motion along a single direction (e.g. horizontal), to satisfy equation 3.

3.2 Brightness Constancy Constraint

For the implementation of Optical Flow, we assume that the intensity is preserved, so we have to ensure the brightness constancy constraint between the stereo pair. To enforce the constraint we match the brightness mean of both images. Specifically, we match the mean of one image regarding the other. In the example presented in Figure 1 we can see that the brightness of both images differs. Brightness equalization is the starting point for Optical Flow estimation. This can be done by adding some value to the intensity of one image so its mean matches that of the other image. Another way could be to ensure that the mean and the variance of both images are equals by multiplying/adding some factor to the intensity.

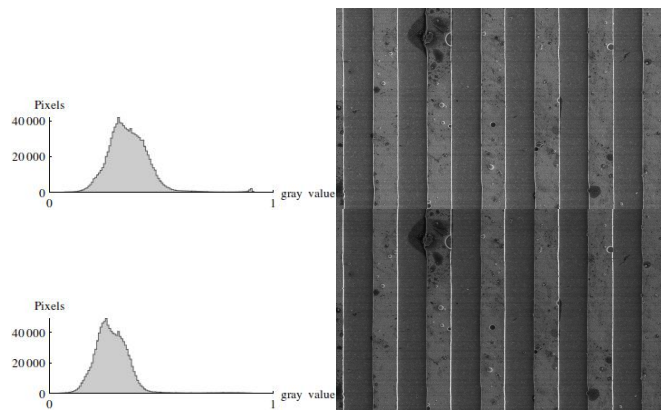


Fig. 1: Comparison between histogram distribution of the two stereo images shown on the right side before intensity correction for brightness constancy.

3.3 Alignment and Geometric Transformation

There is uncertainty during the acquisition process that can produce image misalignments in the stereo pairs. Those misalignments introduce an undesired component in Optical Flow calculation. Aiming to correct such misalignments, a geometric transformation that aligns the stereo pair is approximated.

Firstly, the SURF (Speeded Up Robust Feature) algorithm [10] is used in order to find landmarks in the images¹. The landmarks detected in the two images are matched, providing a set of corresponding points. With the point pairs obtained with SURF, a geometric transformation (Figure 2) is computed. Since the resulting system of equations is overdetermined with erroneous detections and matches, RANSAC (Random Sampling and Consensus) algorithm is used to reject outliers. The resulting transformation matrix

¹SURF is a scale- and rotation-invariant detector and descriptor. The descriptor is based on Haar wavelets and makes an efficient use of integral images [5], [10]

maps the largest number of points from one image to their correspondences in the other image.

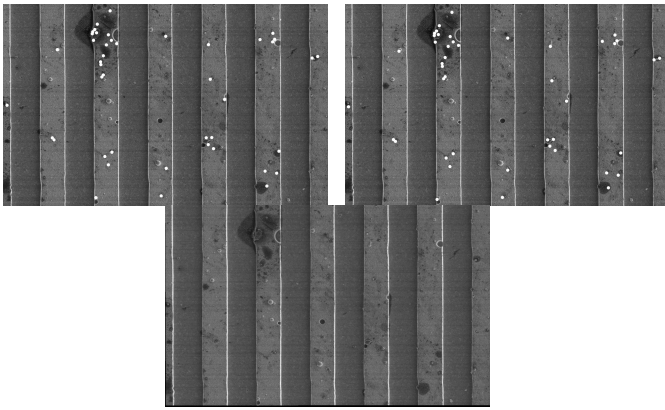


Fig. 2: SURF points of the two stereo images (top) and the resulting aligned image (bottom).

3.4 Optical Flow Implementation

We use a Gaussian low-pass filter for extracting the structure of the signal and improve the Signal-to-Noise Ratio (SNR) [9]. With the Gaussian filter we prevent issues regarding indetermination of the derivatives at the edges. We use a large rotationally symmetric Gaussian low-pass filter with a standard deviation $SD = 25$.

To calculate the derivative with respect to X in equation 3, we applied a 3×3 Sobel filter in the X direction to each image. This filter acts as a discrete differentiation operator [11]. The time (t) derivative is simply the difference between the two images.

Derivative components regarding X and t can be calculated using the filtered images, with equation 3, and therefore velocity can be obtained. Local velocity can be used to find disparities between the two stereo images, which in turn are directly related to the 3D surface relief [5].

4. Experimental Results

After the calculation of the expression in equation 3, we obtain Δx values at each pixel of the images. These values (disparities) are inversely proportional to the height(or depth) of the surface under observation. Therefore the relief of the surface is $k/\Delta x$, where k is a constant depending on the rotation angle, magnification and pixel size [5]. In this section we show the results as a profile of the values, and as a surface mesh.

We used stereo pair images taken from a microscopic grid, a corneal *stromal* surface and a wear metallic surface to test our methodology.

4.1 Profile and Mesh

The profile (Fig. 3) of a sample is obtained by plotting height values contained in one of its image rows. In this profile we can see an approximation of a slice through the 3D surface of the specimen. Also, we made an average of all the rows to view the tendency of the sample. With this average the noise is reduced and the profile is more defined. The profile is useful in samples with a regular pattern, or with a regular roughness or wear, like the microscopic grid or the wear metallic surface. That is because one can obtain a general idea of the whole surface. In samples like the corneal *stromal*, the profile does not give valuable information because of the randomness of the surface, in which one must estimate features of a different kind.

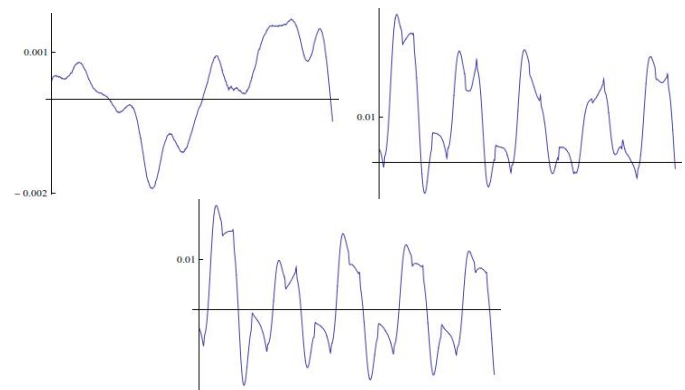


Fig. 3: Wear metallic surface mean profile (top left), Grid surface profile for the middle row (top right) and Grid surface mean profile (bottom).

We visualize with a surface mesh (Fig. 4) the distribution of heights and depths. The values of the optical flow matrix obtained are fundamental for the analysis of characteristics, and the determination of different kinds of properties of the surfaces. On the mesh we can see the areas where the samples are homogeneous and with a constant height, and can confirm the real variation between zones. This information is really important for calculating parameters like roughness and wear in the surface of the samples and to obtain accurate results in their measure.

5. Conclusions

The methodology presented in this paper for obtaining a 3D reconstruction of SEM images using the differential technique of Optical Flow and SEM stereo pairs is fully automated. The results give valuable 3D information of the surfaces of the samples and permit the calculation of different parameters. Despite the fact that SEM acquires three-dimensional-like images, for further analysis of the

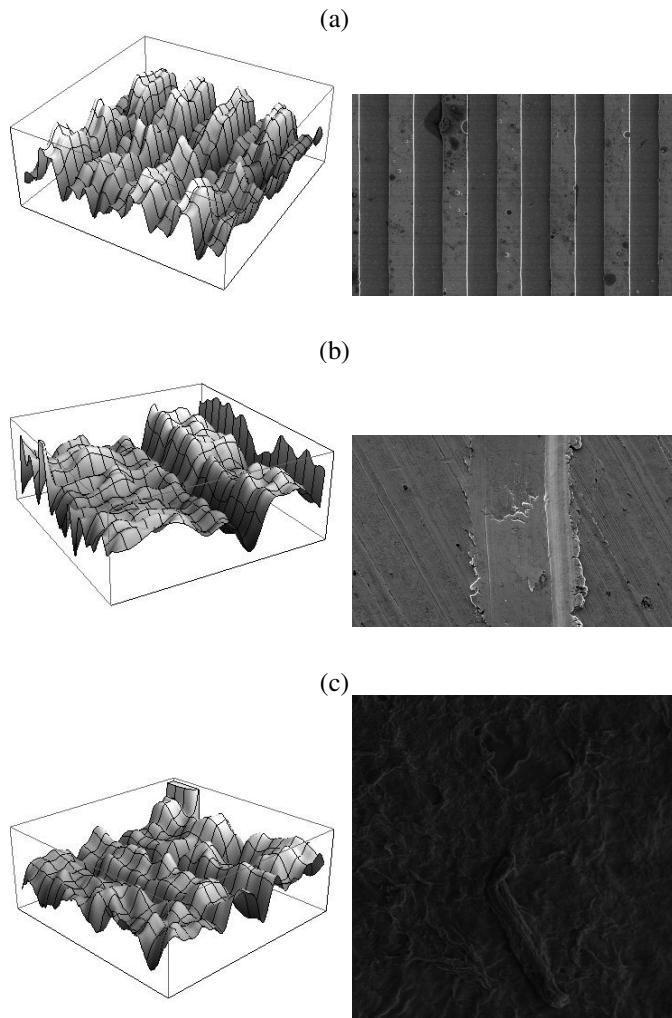


Fig. 4: Surface grid (a), wear metallic (b) and cornea (c) 3D relief reconstruction.

surface, 3D stereo reconstruction provides complementary information, and is the starting point for roughness and wear calculation. With this automated methodology anybody who knows how to use a SEM can obtain an approximation of the true 3D surface of a sample.

It is fundamental to improve the acquisition of the stereo pair to ensure good performance using the proposed methodology. Additional algorithms to reduce the need to take long time obtaining good images in the microscope will also be investigated.

The results of this 3D surface reconstruction are motivating, and as a first approximation are really interesting. With the methodology using optical flow, the surface reconstruction is accurate in comparison with the real shape.

In the future we intend to calibrate the algorithm to make accurate absolute measurements of the 3D reconstructions, and with these data, study the roughness and

wear of different samples from SEM images, specifically the measurements of wear and roughness of thin films.

Acknowledgment

This research was supported by the Ministry of Foreign Affairs, Trade and Development of the Government of Canada, with the scholarship “Emerging Leaders in the Americas Program” (ELAP) and the Administrative Department of Science, Technology, and Innovation - Colciencias with the program “Jóvenes Investigadores e Innovadores 2012” cooperation agreement No. 0729-2012

References

- [1] A. Bogner, P.-H. Jouneau, G. Thollet, D. Basset, C. Gauthier. 2007. A history of scanning electron microscopy developments: Towards “wet-STEM” imaging. *Micron* 38, 390 - 401.
- [2] K. D. Vernon-Parry. 2000. Scanning Electron Microscopy: an introduction. III-Vs Review, Vol. 13 No. 4.
- [3] Joseph I. Goldstein et al. 2003. Scanning Electron Microscopy and X-ray Microanalysis. 3rd Edition. Kluwer Academic/Plenum Publishers, New York.
- [4] Philips Electron Optics. 1996. Environmental Scanning Electron Microscopy. 2nd ed. Robert Johnson Associates, El Dorado Hills.
- [5] S. Roy, J. Meunier, A. M. Marian, F. Vidal, I. Brunette, S. Constantino. 2012. Automatic 3D reconstruction of quasi-planar stereo Scanning Electron Microscopy (SEM) images. 34th Annual International Conference of the IEEE EMBS, San Diego, California, USA.
- [6] Ruggero Pintus, Simona Podda, Massimo Vanzi. 2006. An Automatic Alignment Procedure for a 4-Source Photometric Stereo Technique applied to Scanning Electron Microscopy. IMTC-Instrumentation and Measurement, Technology Conference. Sorrento, Italy.
- [7] S. S. Beauchemin, J. L. Barron. 1995. The Computation of Optical Flow. *ACM Computing Surveys*, Vol 27, No. 3.
- [8] David J. Fleet, Yair Weiss. Optical Flow Estimation. 2005. *Mathematical Models in Computer Vision: The Handbook*, Chapter 15, Springer, pp. 239-258
- [9] J. L. Barron, D.J. Fleet, S. S. Beauchemin. 1994. Performance of Optical Flow Techniques. *IJCV* 12:1, pp43-77.
- [10] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool. Speeded-Up Robust Features. *Computer Vision and Image Understanding* 110 (2008), 346 - 359.
- [11] Samta Gupta, Susmita Ghosh Mazumdar. Sobel Edge Detection Algorithm. *International Journal of Computer Science and Management Research*. Vol 2 Issue 2 February 2013, ISSN 2278-733X.

