OCR for Unreadable Damaged Characters on PCBs Using Principal Component Analysis and Bayesian Discriminant Functions

Carlos F. Nava-Dueñas, *member IEEE*
Skyworks Solutions, Inc.; Engineering Institute
UABC, Mexicali, Mexico
nava.carlos@uabc.edu.mx

Felix F. Gonzalez-Navarro, *member SMIA*
Engineering Institute
UABC, Mexicali, Mexico
fernando.gonzalez@uabc.edu.mx

***Abstract*— In the last few decades, new computer vision technologies and image processing techniques have been very important in the improvement and automation of manual processes in many technical areas, e.g., in the semiconductor industry. In this paper, we propose to change the actual pattern matching methods implemented to have optical character recognition by the use of principal component analysis method to extract the principal characteristics and features of damaged or unreadable numerical digit characters from images on printed board circuits (PCBs) and compute linear and quadratic Bayesian discriminant functions to classify and find the correct numerical character that corresponds to those features. In the first step of this work, grayscale color images are acquired from a charge-coupled device (CCD) camera, then image segmentation is manually computed to create a dataset of 500 matrix images for the character digits from 0 to 9. Then, a feature extraction method is applied to get the principal components that will be used in the character recognition state. Finally, our results show that applying Bayesian linear and quadratic discriminants to the principal component features can improve optical character recognition (OCR) detectability of damaged characters from actual 95–97% to 99.88% in early tests. This suggests to us that the problem probably follows a linear model where linear hyperplanes separate decision regions with satisfactory (almost no) errors.**

Keywords*— *machine vision; optical character recognition; Bayesian discriminant functions; principal component analysis; linear discriminant classifier; quadratic discriminant classifier.

I. INTRODUCTION

Optical character recognition (OCR) has been an important technology used to convert characters from a digital image to a digital text. There are basically two types of OCR algorithms: the first technique is related with the matching of matrix images, where an alphabet of stored character images is used to compare with an input image [1], [2]. This pattern matching does not work well when new fonts are encountered or input character images are unreadable. The second technique decomposes an input image to extract the principal features [3], [4], [5]. Then, classifiers are used to compare the input image features with some stored image features and choose the best match.

Our actual system implemented at the Skyworks factory uses the traditional OCR technique, pattern matching. Our implemented vision system reads identification characters on printed circuits boards (PCBs) for lot integrity and machine control. This commonly used technique is not robust enough because many of the images of PCBs shown some damage on the characters due dirt or the results of bad previous processes [1], [6]. Our actual OCR detectability is around 97% at best. Our system starts with a monochrome VGA image acquisition of the upper left section of a PCB, using a NI-1752 smart camera, with a full resolution of 640x480 pixels with a maximum data transfer of 60 fps using a GigE port. The selected resolution and data transfer speed parameters meet the factory production schedule of inspected PCBs. The camera has a grayscale output image type with a maximum character resolution to cover the entire PCB characters positions, as shown in Fig. 1.

Fig. 1. PCB with no damaged characters.

As mentioned before, due to problems with previous processes in the production line, some PCBs present some residual dirt over the characters, making some characters unreadable for the pattern matching technique, as shown in the following Fig. 2.

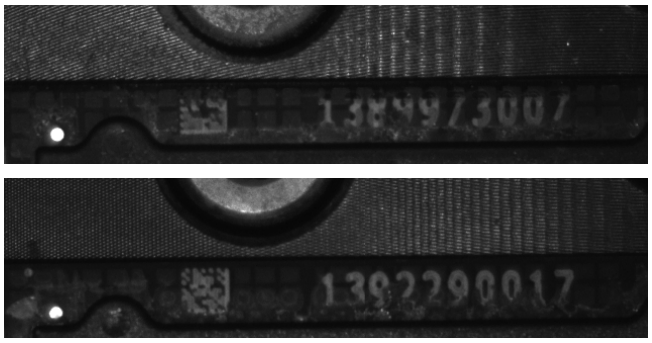

Fig. 2. PCBs with evident residual dirt over characters.

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.165

The principal problem is that operators have lower throughput than automatic OCR software, and this leads to manually writing down the information from the screen when the actual recognition software fails, increasing the process time, making possible errors from wrong readings, inducing higher production costs. Taking into consideration these facts, a better approach has to be considered [3].

This paper presents the proposal for implementing a character recognition technique for unreadable characters using extraction features and Bayesian classifiers.

II. Data Set Constructor

Our implementation starts with an experimental dataset constructed of 500 character images. In this dataset, we have 50 images that correspond to each numerical digit image from 0 to 9. Next, Fig. 3 shows several digit image samples.

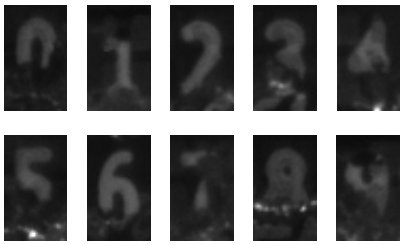

Fig. 3. Some damaged digit images from dataset

For our previous dataset, we let I_i any (k,l) digit image, Fig.4, from the original dataset. $\forall I_i$:

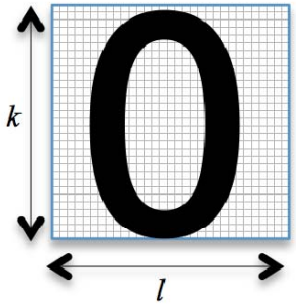

Fig. 4. I_i digit image matrix with size (k,l)

- Convert I_i to gray-scale (if previous images are RGB type).
- Transform matrix I_i to a row-vector of size (1, k*l).
- Create a matrix M of size (n, k*l), where n is the number of samples for each digit image, $M \leftarrow M \cup I_i$.

III. Principal Components Analysis and Bayesian Classification

The purpose of this portion of the paper is to map the matrix M into the eigenspace by means of the the first P principal components. We follow the next steps:

1) Extract the mean for each column:

$$\Phi_j \leftarrow \frac{1}{n}\sum_{i=1}^{n} m_{ij}$$
$$\hat{M} \leftarrow M_j - \Phi_j \quad (1)$$

2) Compute tne covarıance matrıx γ of $\widehat{M}$.

3) Compute eigenvectors and eigenvalues (PC,V) of γ.

4) Sort matrix PC by columns in descend order ruled by vector V.

5) Project $\widehat{M}$ into the first P principal components:

$$X \leftarrow \hat{M} * PC \quad (2)$$

6) A new digit dataset is now assembled, X.

The principle component analysis (PCA) digit image test dataset must be processed by applying the first 5 steps, but using Φ_j and $\widehat{M}$ calculated from the training set.

Given the new Eigen-data set, two Bayesian algorithms, linear and quadratic discriminant classifiers must be trained and tested by means of 10x10 cross-validation method. These algorithms are widely used parametric methods, which assume that the class distributions are multivariate Gaussian [7], [8], [9].

With linear discriminant analysis (LDA), all classes are assumed to have the same covariance matrix, but quadratic discriminant analysis (QDA) does not need such an assumption; however, the number of parameters to be estimated from the data available for each class is much higher, entailing lower statistical significance. The discriminant functions associated to each classifier are defined as:

1) Linear Discriminant Classifier:

$$g_k(\boldsymbol{x}) = \ln P(\omega_k) + \boldsymbol{\mu}_k^t \Sigma^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^t \Sigma^{-1} \boldsymbol{\mu}_k \quad (3)$$

2) Quadratic Discriminant Classifier:

$$g_k(\boldsymbol{x}) = \ln P(\omega_k) - \frac{1}{2}\left(\ln |\Sigma_k| + (\boldsymbol{x}-\boldsymbol{\mu}_k)^t \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)\right) \quad (4)$$

IV. Experimental Results

In the following Fig. 5 we show the first 3 principal components from matrix *X* with only 3 characters of data: 0, 1 and 2.

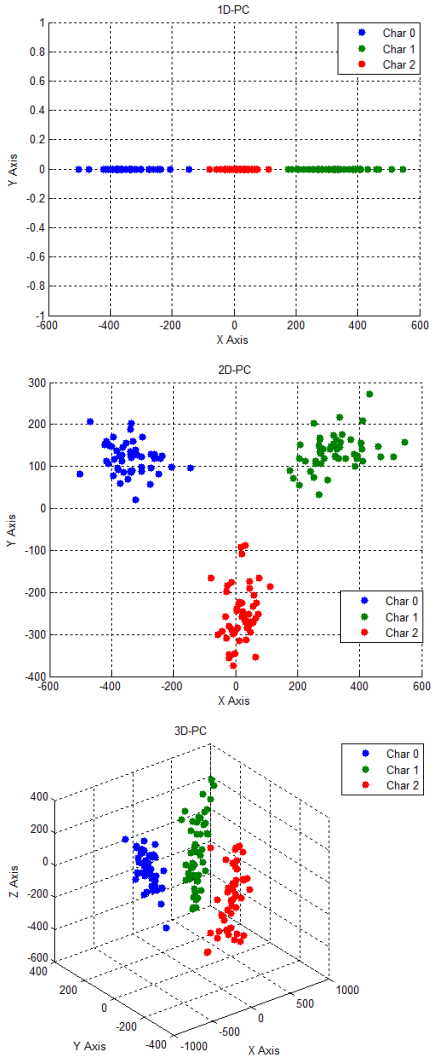

Fig. 5. 1 PC, 2 PC and 3 PC for characters 0, 1 and 2.

For more digit characters, the principal components were not easy to visually classify, as shown in Fig. 6:

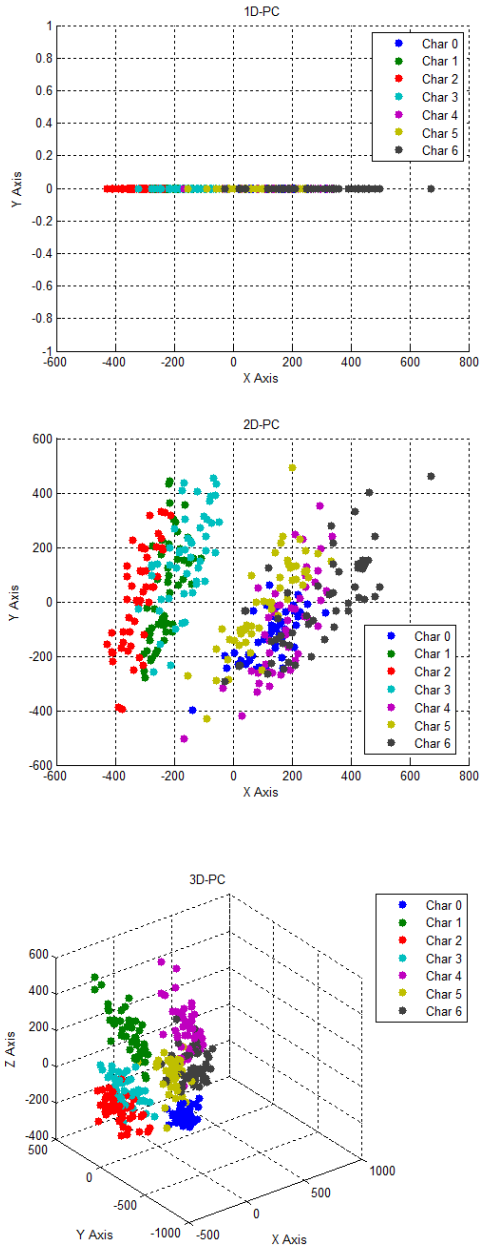

Fig. 6. 1 PC, 2 PC and 3 PC for Characters 0 to 6.

From previous simulations we can see that using 2 or 3 principal components is not enough to have a difference in the proximity of the characters groups. It's clear that groups for characters 0, 1 and 2 are close. The next step will use the linear and quadratic discriminant classifiers using more than 3 principal components from *X* matrix.

Our classification process, for linear and quadratic discriminants, was trained in the complete training data set and tests the performance in the test data set. A 10x10 cross-validation model validation technique was computed to estimate how our classification model was performed [10]. Fig.7 shows these experimental results:

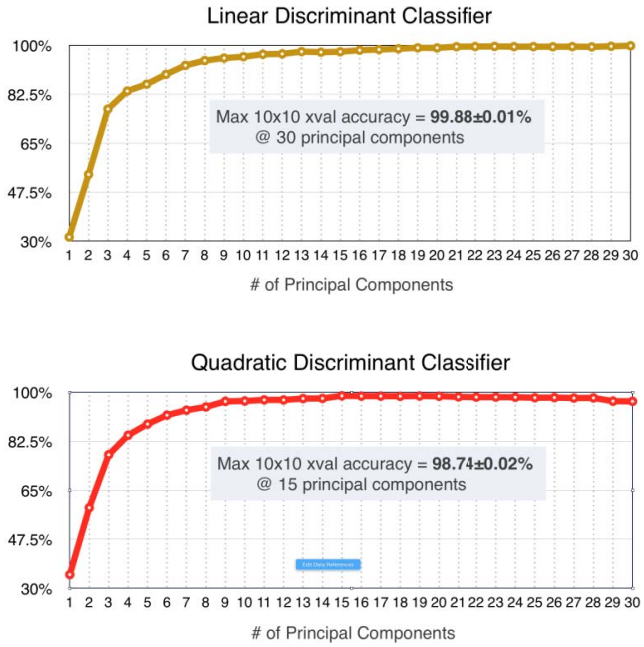

Fig. 7. 10x10 cross-validation for LDC and QDC classifiers.

For previous simulations, it's seen that classification algorithms yield promising results. The 10x10 cross-validation recognition rate for the linear discriminant classifier shows an interesting 99.88%, by using the first 30 principal components. To check the performance of the linear discriminant classifier we compute the confusion matrix, where it shows a 100% of recognition rate at characters 0, 1, 3, 6, 7, 8 and 9. Relative difficulties are seen in character 2, which was misclassified as character 7 in one case. Character 4 was classified as character 1 in just one case.

Quadratic discriminant classifier performance was as follow: 10x10 cross-validation recognition rate of 98.74% by means of the first 15 principal components. The average confusion matrix shows perfect recognition in only one character, 9.

V. Conclusions

The implementation of principal components and Bayesian linear and quadratic discriminants demonstrates an improvement for the readings in optical character recognition from a previous detectability from pattern matching of 97% — at best to close to 99% — for this paper technique. In this paper we can conclude that our data follows almost a linear nature. However, this assertion should be taken with caution, given that the data comes from one machine. Future work will increase the dataset to more machines and other classification techniques will be included, like artificial neural networks.

References

[1] S. Mori, C. Suen and K. Yamamoto, "Historical review of OCR research and development", *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1029-1058, 1992.

[2] R. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 690-706, 1996.

[3] Miciak, M., "Character Recognition Using Radon Transformation and Principal Component Analysis in Postal Applications", *Proceedings of the International Multiconference on Computer Science and Information Technology*, p. 495-500, 2008.

[4] I. T. Jolliffe, "Principal Component Analysis", *Springer Series in Statistics*, 2nd ed., Springer, 2002.

[5] S. Nedevschi, I. Peter and A. Mandrut, "PCA type algorithm applied in face recognition", *2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing*, 2012.

[6] D. Desrochers, Z. Qu and A. Saengdeejing, "OCR readability study and algorithms for testing partially damaged characters", *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*, 2001.

[7] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.

[8] I. Ibraheem, "Linear and Quadratic Classifier to Detection of Skin Lesions "Epicutaneus"", *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, 2011.

[9] Yang, Y., Qiu, Y., and Lu, C., "Automatic target classification & experiments on the MSTAR SAR images", In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks* (SNPD/SAWN 2005), May 23—25, 2005, 2—7.

[10] Ron Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the 14th international joint conference on Artificial intelligence,* p.1137-1143, August 20-25, 1995, Montreal, Quebec, Canada.