DT-Binarize: A Hybrid Binarization Method using Decision Tree for Protein Crystallization Images

İmren Dinç[1], Semih Dinç[1], Madhav Sigdel[1], Madhu S. Sigdel[1], Marc L. Pusey[2], Ramazan S. Aygün[1]
[1]DataMedia Research Lab, Computer Science Department
University of Alabama in Huntsville
Huntsville, Alabama 35899
[2]iXpressGenes Inc., 601 Genome Way, Huntsville AL, 35806
Email: {id0002, sd0016, ms0023, mss0025, aygunr} @uah.edu[1], marc.pusey@ixpressgenes.com[2]

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, 90^{th} 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, likelyleads, 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.

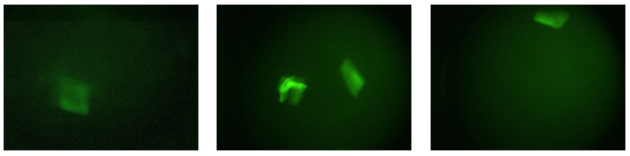
2. Background

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.

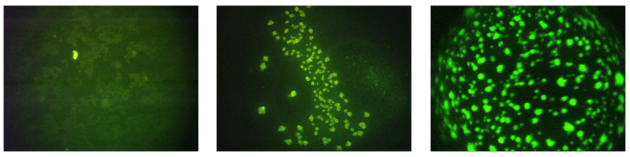


(a) (b) (c)

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.

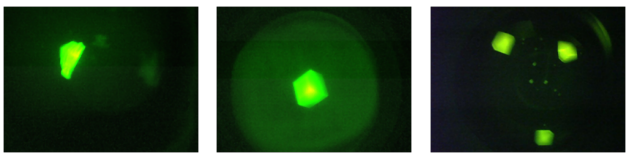


(a) (b) (c)

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.



(a) (b) (c)

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 Otsuś thresholding [5], firstly a gray level image is generated from an input color image. Then, for each possible intensity threshold, the varince 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 90^{th} 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 90^{th} 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 90^{th} 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.

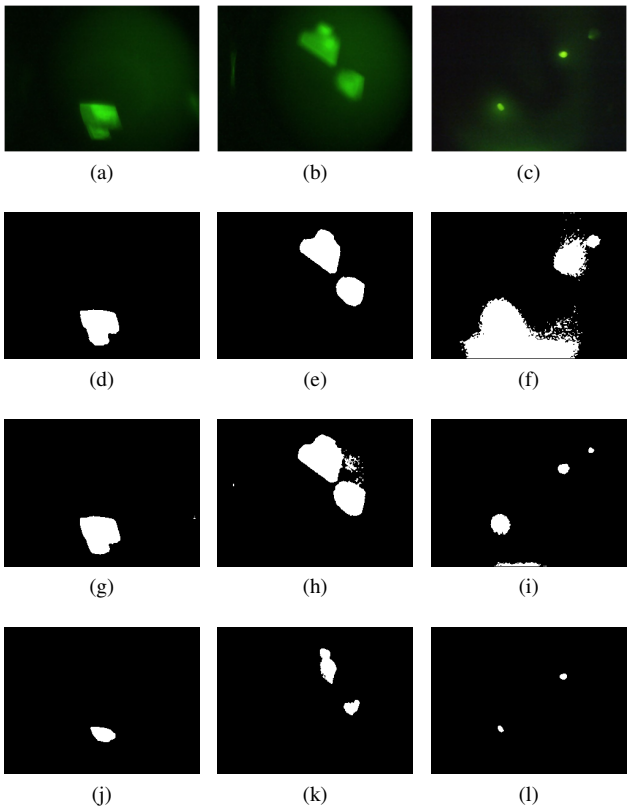


Fig. 4
EXAMPLE IMAGES THAT WORKS GOOD FOR ALL THRESHOLDING TECHNIQUES, (A), (B), (C) ORIGINAL IMAGES, (D), (E), (F) OTSU RESULTS, (G), (H), (I) G90 RESULTS, (J), (K), (L) G100 RESULTS, NOTE THAT (E),(G), AND (L) ARE THE BEST BINARY IMAGES

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

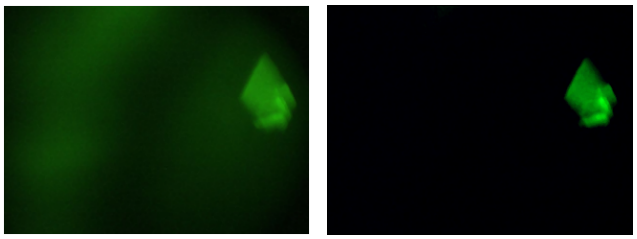
Contrast stretching is a normalization method that enhances the informative features of the image by expanding the histogram of the intensities. It maps the pixel values into a new range in a linear fashion [6]. We can apply contrast stretching to the images by using the Eq 1,

$$I_{out} = (I_{in} - P_{in})(\frac{P'_{max} - P'_{min}}{P_{max} - P_{min}}) + P'_{min} \qquad (1)$$

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 loosing 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.



(a) (b)

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.

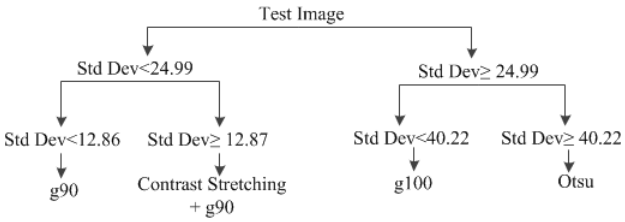


Fig. 6

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,

$$I_S = 2 \times I_R + I_O \tag{2}$$

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.

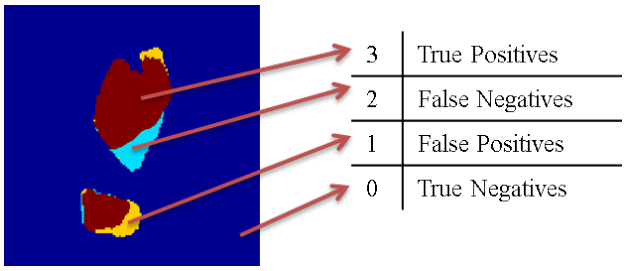


Fig. 7
EXAMPLE SUM IMAGE

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.

$$M_{acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$M_{F1} = \frac{2 \times TP}{(2 \times TP) + FP + FN} \tag{4}$$

$$M_{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

$$M_{Jacc} = \frac{TP}{TP + FP + FN} \tag{6}$$

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

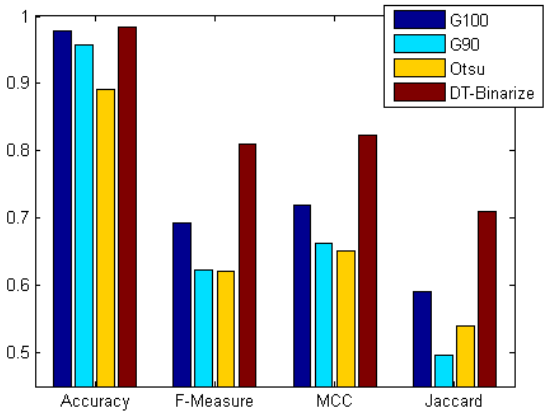


Fig. 8
COMPARISON RESULTS

our technique can successfully generate the best result. Our DT-Binarize method can adapt different lighting and focusing conditions.

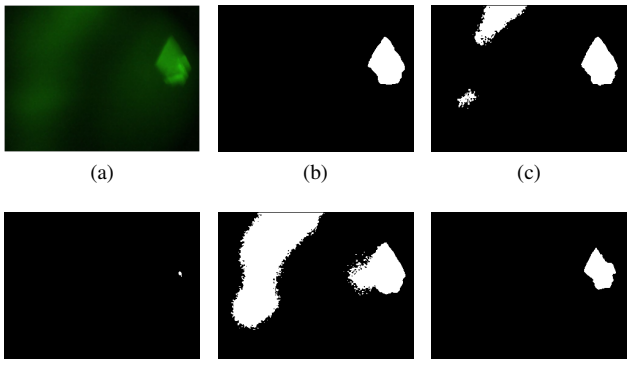


Fig. 9
A SAMPLE TEST CASE (A) ORIGINAL IMAGE, (B) GROUND TRUTH IMAGE, (C) g_{90} THRESHOLD, (D) g_{100} THRESHOLD, (E) OTSU THRESHOLD, (F) DT-BINARIZE

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.

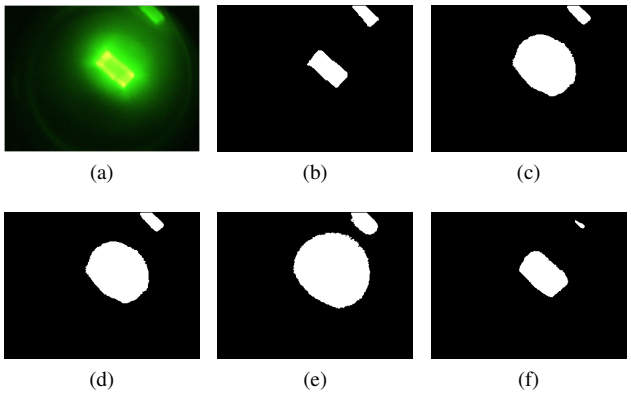


Fig. 10
EXAMPLE OF A BAD BINARIZATION RESULT (A) ORIGINAL IMAGE, (B) GROUND TRUTH IMAGE, (C) g_{90} THRESHOLD, (D) g_{100} THRESHOLD, (E)OTSU THRESHOLD, (F)DT-BINARIZE

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.

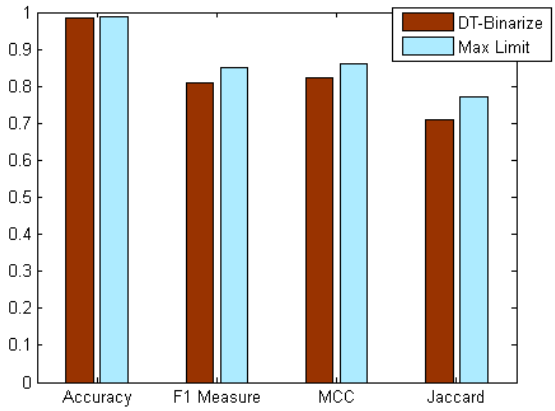


Fig. 11
COMPARISON WITH THEORETICAL MAX LIMIT

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 or 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