Image-Based Classification for Automating Protein Crystal Identification

Xi Yang[1], Weidong Chen[1], Yuan F. Zheng[1, 2], and Tao Jiang[3]

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China
[2] Electrical & Computer Engineering, The Ohio State University, USA
[3] National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Science, Beijing 100101, China
wdchen@sjtu.edu.cn

Abstract. A technology for automatic evaluation of images from protein crystallization trials is presented in this paper. In order to minimize the interference posed by the environmental factors, the droplet is segmented from the entire image first. The algorithm selects different features, which are derived from the pixels within the droplet, and obtains a 16-dimensional feature vector which will then be fed to the classifier to make a classification. Each image is classified into one of the following classes: “Clear”, “Precipitate” and “Crystal”. We have achieved an accuracy rate of 84.8% with our algorithm.

1 Introduction

The analysis of the protein structure is an important component of protein crystallography, which has been one of the most popular research areas in recent years. Study of the function of protein crystal helps us to understand the mechanism of the protein as well as the interplay between protein molecule and other molecules [1].

The high-throughput protein crystallization system can prepare thousands of trials per day. Conventionally, the outcomes of the protein crystallization trials are assessed by human experts. This procedure is slow and inefficient. Therefore, an automatic technology needs to be studied to replace the manual work.

Several methods have been proposed by other researchers [2 - 5]. The best result was achieved by Bern et al. in 2004 [6]. However, when their algorithm is applied to our image set, the accuracy rate is not acceptable.

In this paper, we propose an automatic protein crystallization classification algorithm, which is based on the digital image processing technology. All the image samples obtained from the protein crystallization equipment are classified into 3 different classes, called “Clear” – no substance is produced, “Precipitate” – the primary substances produced are precipitates, and “Crystal” – the primary substances produced are crystals.

2 Methodology

The procedure of the algorithm, as shown in Fig. 1, consists of 3 steps, including image segmentation, feature extraction and classification. Otsu automatic threshold [7],

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNCIS 345, pp. 932 – 937, 2006.

Canny edge detection [8] and Active Contour Model (ACM) [9] are utilized to locate the boundary of the droplet. Image features are derived by calculating the Gray Level Co-occurrence Matrix (GLCM) [10], Hough Transform and Discrete Fourier Transform (DFT) of all the pixels belonging to the droplet. The classification procedure is divided into two stages; each of which is a two-class problem.

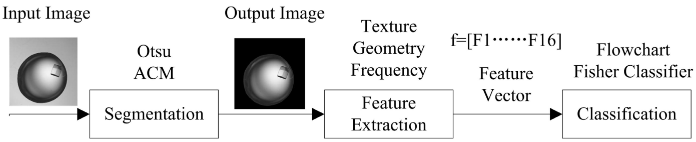

Fig. 1. Procedure of the algorithm

3 Algorithm

3.1 Image Segmentation

In our algorithm, firstly, the image is converted from gray scale to binary image. The threshold used is computed by the Otsu algorithm.

Then an ACM is applied to the image. ACM is defined as a dynamic contour which can change its shape, based on its energy function, to adapt to the local feature, for example, the boundary of the droplet. The energy function of ACM can be expressed as equation (1):

$$E = E_{\text{int}} + E_{ext} \tag{1}$$

E_{int} is the internal energy based on the shape of the dynamic contour, and E_{ext} is the external energy based on the local feature of the image. Interested readers are referred to [9] for the extra formulas. The internal energy will smooth the dynamic contour and the external energy will adapt it to the edges detected in the image when minimizing the energy function.

The ACM should be initialized with a position, from which it begins to change its shape. The minimum circle surrounding the connected component, which has the maximum area, is selected as the initial location of the contour. Canny edge detection is employed to detect the edge of the droplet, which is taken as the final position of the dynamic contour. After several iterations, the contour will converge to the boundary of the droplet, and all the pixels within the contour can be segmented from the image. The procedure of image segmentation is shown in Fig. 2.

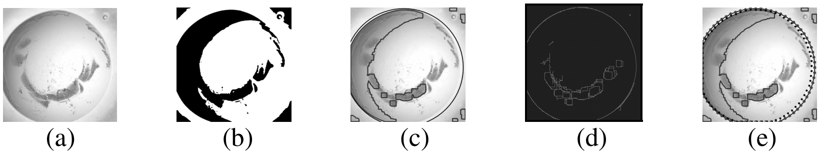
(a) (b) (c) (d) (e)

Fig. 2. Procedure of the segmentation. (a): original image; (b): binary image converted from (a); (c): initial circle; (d): image with edges detected; (e): the nodes represent the boundary detected by ACM.

3.2 Feature Extraction

The classification is made based on the features extracted from the image. At first, individual features are extracted from the image, and then a feature vector is formed by these individual features. The features used include the texture, geometry and frequency features derived from the pixels inside the droplet. A detailed description of the features is presented as follows:

A notable characteristic of the image is the texture features based on the statistical analysis of the image. The images with and without crystals will present different texture features in terms of contrast, correlation, etc. Texture features are obtained by computing the GLCM of the pixels belonging to the droplet. The GLCM is defined by the following equation:

$$p(i,j)=\#\left\{(x,y)\,|\,f(x,y)=i,and\begin{pmatrix}f(x+\Delta x,y+\Delta y)=j\\ or\\ f(x-\Delta x,y-\Delta y)=j\end{pmatrix},x,y=0,1.....\mathrm{N}-1\right\} \tag{2}$$

$p(i,j)$ is the element of the GLCM. x and y are the coordinates of each pixel, and $f(x,y)$ represents the gray scale value of that pixel. $\#\{\Omega\}$ means the number of elements within the set. Four properties can be computed from the GLCM: Entropy, Energy, Contrast and Correlation, which are selected as the texture features together with the mean value and the standard deviation of the gray scale values of all the pixels within the droplet.

Another significant property of the image is the straight lines detected in the droplet. It can be seen that the edges of the crystals are always presented as straight lines while the edges of precipitates are usually curves. Hough Transform is utilized to detect the straight lines. Two values are taken as the geometry features, as ever mentioned by Cumbaa et al. [5], the total length and the maximum length of the lines detected in the image.

Note that for the protein crystals, their edges are always clear and sharp while for the precipitates, their edges are always fluffy and smooth. As a result when the images are converted from the spatial to frequency domain, the images with precipitates will have more energy at high frequency components than the images with crystals. We select the mean values and the standard deviations of the image energy at four different frequency bands as the frequency features.

3.3 Classification

The classification is formed by two steps. In the first step, each image needs to be classified into “Clear” – no substance is generated or “Not Clear” – something (either precipitates or crystals) is generated. In the second step, the images which are labeled as “Not Clear” are classified into “Precipitate” and “Crystal”.

The parameters used in the first step are defined as follows:

A1: the number of grids, whose gray scale standard deviation exceed TC1.
A2: the number of grids, whose gray scale standard deviation exceed TC2.
A3: the number of grids, whose entropy exceed TC3.
K: The number of connected components detected within the binary image.
TC1, TC2 and TC3, TA1, TA2 and TA3 are manually determined thresholds.

Firstly, the image is divided into several 30×30 pixels grids. If the A1 value of the image exceeds TA1, the image is marked as “Not Clear”. Otherwise, the image is converted from gray scale to binary image by the self-adaptive threshold algorithm. We detect the connected components within the binary image, and compute K. If K is not zero, then draw several rectangles surrounding each connected component, and divide each rectangle into 5×5 grids. If the A2 value of either rectangle exceeds TA2, the image is “Not Clear”. If K is zero or A2 is smaller than TA2, again, divide the segmented image into several 30×30 pixels grids. If the A3 value of the image is greater than TA3, then the image is “Not Clear”. Otherwise, the image is “Clear”. The flowchart shown in Fig. 3 describes the algorithm performed in the first step.

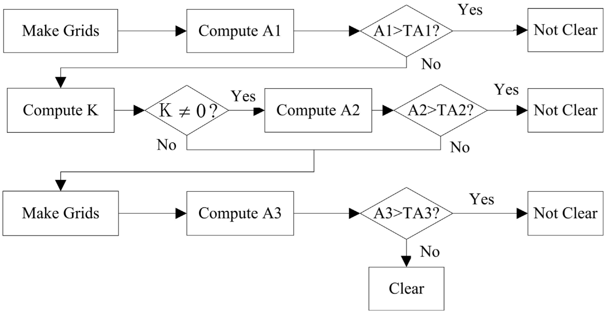

Fig. 3. The flowchart of the algorithm used in the first step of the classification

In the second step, a Fisher classifier, which has been trained by a human labeled learning set, is employed to make the classification. The learning set consists of images with crystals, called positive samples, and images with precipitates, called negative samples. The 16-dimensional feature vector f can be obtained from each image. Compute a project vector w, which should satisfy the following condition: when each f is projected onto w, the positive and the negative samples should be maximally separated. For an image with unknown class, compute its feature vector f, if $f \cdot w$ exceeds a scalar quantity l, then label the image as “Crystal”. Otherwise, label the image as “Precipitate” as shown in equation (3). The scalar quantity l can be determined by prior knowledge.

$$f \cdot w \begin{cases} \geq l \rightarrow Crystal \\ < l \rightarrow Precipitate \end{cases} \tag{3}$$

4 Experimental Results

The learning set is formed by 10 images with crystals, and 10 images with precipitates. The testing set comes from a combination of 52 “Clear” images, 12 “Precipitate” images and 46 “Crystal” images. The experiment is performed on a PC with Windows XP operating system, and the CPU is AMD 2500+. With the project vector w derived from the learning set, we achieve a result as shown in Table 1:

Table 1. Result of the experiment

True	Detected		
	"Clear"	"Precipitate"	"Crystal"
"Clear" (52)	**82.7% (43)**	1.9% (1)	15.3% (8)
"Precipitate" (12)	8.3% (1)	**58.3% (7)**	33.3% (4)
"Crystal" (46)	2.2% (1)	13.0% (6)	**84.8% (39)**

Typical images processed in the experiment are shown in Fig. 4.

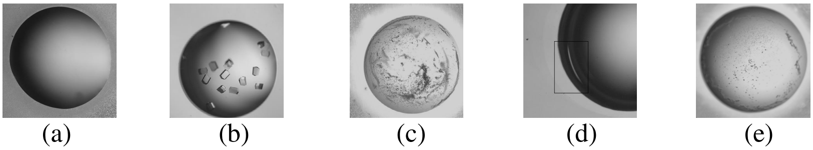

Fig. 4. Typical images processed in the experiment. (a) (b) and (c) can be classified correctly, where (a) is "Clear", (b) is "Crystal", and (c) is "Precipitate"; (d): "Clear" image is classified as "Crystal" due to the light reflection as shown in the block; (e): image which contains grainy crystals is falsely classified as "Precipitate".

5 Conclusion

The algorithm proposed in this paper is proved to be effective and efficient. 84.8% "Crystal" images can be recognized correctly in the experiment. In order to increase the accuracy rate, new features should be considered, for example, the ones suggested by Bern et al. [6], corners, transparency and closed outer contours. Besides DFT used in our algorithm, wavelet transform can also be utilized to obtain more information.

Finally, although the images with crystals can be differentiated from those with precipitates, the capability and quality of each protein crystallization trial are still unknown. The number of the crystals generated and the size of each crystal need to be studied to evaluate the performance of each trial in the future.

Acknowledgement

This work is partly supported by the National Hi-Tech Research and Development Program under grant 2005AA420010.

References

1. Abola, E., Kuhn, P., Earnest, T., Stevens, R.: Automation of X-ray Crystallography. Nature Structural Biology, 7 (2000) 973-977
2. Wilson, J.: Towards The Automatic Evaluation of Crystallization Trials. Acta Crystallographica D, vol. 58 (2002) 1907-1914

3. Spraggon, G., Lesley, S. A., Kreusch, A., Prestle, J. P.: Computational Analysis of Crystallization Trials. Acta Crystallographica D, vol. 58 (2002) 1915-1923
4. Jurisica, I., Rogers, P., Glasgow, J. I., Fortier, S., Luft, J. R., Woilfley, J.R.: Intelligent Support for Protein Crystal Growth. IBM System Journal, vol. 40, no. 2 (2001) 394-409
5. Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C.: Automatic Classification of Submicrolitre Protein-crystallization Trials in 1536-well Plates. Acta Crystallographica D, vol. 59 (2003) 1619-1627
6. Bern, M., Goldberg, D., Stevence, R. C., Kuhn, P.: Automatic Classification of Protein Crystallization Images Using A Curve-tracking Algorithm. Journal of Applied. Crystallography D, vol. 37 (2004) 279-287
7. Otsu, N. A.: Threshold Selection Method from Gray-level Histograms. IEEE Trans. Systems, Man and Cybernetics, vol. 9, no. 1 (1979) 62-66
8. Canny, J.: A Computational Approach to Edge Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6 (1986) 679-698
9. Kaas, M., Witkins, A., Terzopolus, D.: Snakes-Active Contour Models. International Journal of Computer Vision, vol. 1, no. 4 (1987) 321-330
10. Haralick, R., Shanmugan, K., Dinstein, I.: Textural Features for Image Classification. IEEE Trans. Systems, Man and Cybernetics, vol. SMC-3, no. 6 (1973) 610-621