CS223B Project Interim Report

Classification of Protein Crystallization Imagery

Samuel Cheng, Shaohua Sun and Xiaoqing Zhu
{samc@stanford.edu, shaohuas@stanford.edu, zhuxq@stanford.edu}

Abstract

We investigate the problem of automatic classification of protein crystallization images, which are captured by a robotic setup for protein crystal growth. Building on an initial framework for image processing and classification, we propose to experiment with a number of new methods: the level-set method for boundary detection and segmentation; new feature extraction algorithms based on wavelet and texture filtering; and a support vector machine (SVM) for classification and feature selection. We show initial results after combining the original features with SVM classification, and discuss future plans for an iterative design of the feature extraction and the SVM.

1 Introduction

One of the most challenging problems in structural biology today is to discover the three-dimensional structures of protein molecules. Among the many techniques in use, single-crystal X-ray crystallography, using diffraction, has proven to be most effective and therefore the most popular. A major difficulty in crystallography is the growth of protein crystals, since the outcome is very sensitive to conditions such as chemical solution, temperature, and air pressure. To successfully produce protein crystals suitable for X-ray diffraction, hundreds of thousands of trials may be needed.

A recent solution to the time-consuming and expensive crystallization process has been the use of high-throughput (HT) crystallization robots, which can be set up to output over 100,000 digital crystallization photographs per day[1]. The robots are capable of dispensing small amounts of the reacting agents and protein into a well according to a specified formula, and of recording digital photographs of the crystallization outcomes at fixed time intervals. The automated experimental setup allows a larger number of conditions to be tried for crystal growth. Unfortunately, the amount of data generated by such a system is enormous. Pure manual classification of the crystallization outcome images is thus no longer feasible; automatic classification is needed to aid in final human decisions.

Many research efforts have been devoted to the classification problem, which has turned out to be hard even for humans. Figure 1 illustrates three typical classes of the experiment outcomes: *clear*, *precipitate(PPT)* and *crystal*. PPT refers to failed attempts where precipitates instead of crystals are formed in the solution. Crystal refers to successful growth of protein crystals, which may or may not be suitable for X-ray diffraction. "Clear" images

contain neither precipitates nor crystals. Note that the appearance of PPT outcomes can vary significantly under different conditions (see Fig. 2). In addition, the boundary between crystal and precipitate images is also somewhat blurred when the outcome contains both precipitates and small crystals. In this case one would prefer to classify the image as "crystal" for further human inspection. Consequently, current performance of automatic classification of the outcome images typically has false negatives and false positives on the order of 20% [1][2][3]. Best results are reported in [4] with 12% false negatives and 14% false positives. To facilitate the more ambitious goal of automatic analysis of the conditions for successful crystallization, better performance is needed from the automatic classification process.

The goal of this project is to investigate several new methods of improving the performance of automatic classification. We build our experiments upon the framework described in [5], and substitute each major component with our proposed algorithm: boundary detection and segmentation based on the level-set method; extractions of new features using wavelet and texture filtering; and classification and feature selection with the support vector machine (SVM).

In this interim report we show initial results of applying the SVM to some simple geometry-based features (see [4]). The potential gain by iteratively designing the SVM classifier and the feature extraction process is also discussed. In the final report, we wish to evaluate the performance of the proposed methods against results reported in literature.

2 Background

Modern robotic crystallization systems can carry out more than 10,000 trials per day, each with a unique set of chemical conditions and sequentially recorded outcomes via digital photography. Fast and automatic evaluation of the resultant data, however, is still a challenge [3]. The key issue is to distinguish successful crystallization results from unsuccessful ones, i.e., to classify the outcome images into two or more categories. This is a machine learning problem, and has been addressed by many researchers. In this section we give a brief survey of existing methods for automatic recognition and classification of crystallization imagery.

Early work by Zuk and Ward [2] uses the Hough transform to detect straight edges for crystals, but does not attempt to classify the images. A custom-built image acquisition and image processing system is reported in [1], where Jurisica et al. detected the drop boundary by fitting a conic curve and classified the images using spectral analysis. No specific error rates are reported. Instead, the correlation between extracted features and crystallization results is demonstrated by example.

In [3], Wilson proposes an algorithm to find drop boundaries using the Sobel edge detector and to detect objects with high circularity; he then classified the images into three categories based on features of edge pixels and reported accuracy rate at around 75%. Spraggon et al. have tried the Canny edge detector and circle fitting for drop boundary detection and a self-organizing neural net for classification [6], using features related to straight lines and textures and reporting false negatives and false positives about 25%.significantly One problem with this drop detection algorithm is that it may miss cryastals outside the fitted circle. More recently, a probablistic graphical model is used to find the drop boundary and

(a) (b) (c)

Figure 1: Sample images for three classes: clear, precipitate(PPT) and crystal

(a) (b) (c)

Figure 2: Sample images for different appearance of precipitate images

classfication fetures based on correlation filters and the Radon transform (similar to Hough transform, with polar coordinate parameterization) [7]. A balanced error rate of 15% for the two-class classification: crystal-positive and crystal-negative is achieved.

The best result reported so far is in [4], where Bern et al. propose a line tracking algorithm for drop boundary detection and a decision-tree classifier with hand-crafted thresholds operating on the gradient- and geometry-related features of a selected drop. The drop boundary detection has a success ratio of 93%, whereas the classification achieves 12% false negative and 14% false positive. It is also mentioned that the current feature-detection algorithms fail to capture the difference between swirly precipitates due to convective currents and microcrystals, and that new features representing global characteristics of the images are needed.

3 Proposed Approach

An initial codebase for the system in [4] is provided by Dr. Bern at Palo Alto Research Center (PARC). We plan to investigate evaluate new methods for each individual component. In particular, we wish to try the level-set method [8] for boundary detection, wavelet and texture filters [9][10] for new feature extraction, and the support vector machine for classification [11]. In addition, since results from the support vector machine can indicate which feature vectors carry the differentiating characteristics between the two classes, we propose to iteratively design the features and the parameters in the SVM for improved performance. A block diagram of the proposed system is shown in Fig. 3. For the project, each of the participants will focus on one component of the system.

3.1 Boundary Detection

Drop segmentation is the first step in a classification system. It is important to separate the drop from the background well, so that the position of the droplet in the well is not important to the classifier. In addition, the segmentator must be conservative and not classify any features outside the droplet, as the rarity of crystallinity makes false negatives extremely undesirable.

There are some problems with the currently-implemented techniques for boundary detection. Edge detection techniques must be modified in order to generate closed contours, and circle fitting fails on the 15% of droplets with irregular shape. Dr. Bern's parametric "snake" algorithm handles irregular shapes well, but sometimes follows false curves, segmenting some of the droplet into the background. Jurisica et. al. demonstrate the promise of intrinsic deformable models using a conic section, and hopefully this promise will be realized in our implementation.

Use of the level set algorithm should work on irregular shapes, and also be conservative with respect to background segmentation. [12] demonstrates the effectiveness of level set algorithms in segmenting pollen grains from background in electron micrographs. This is the approach we will use to segment the droplet from the well background.

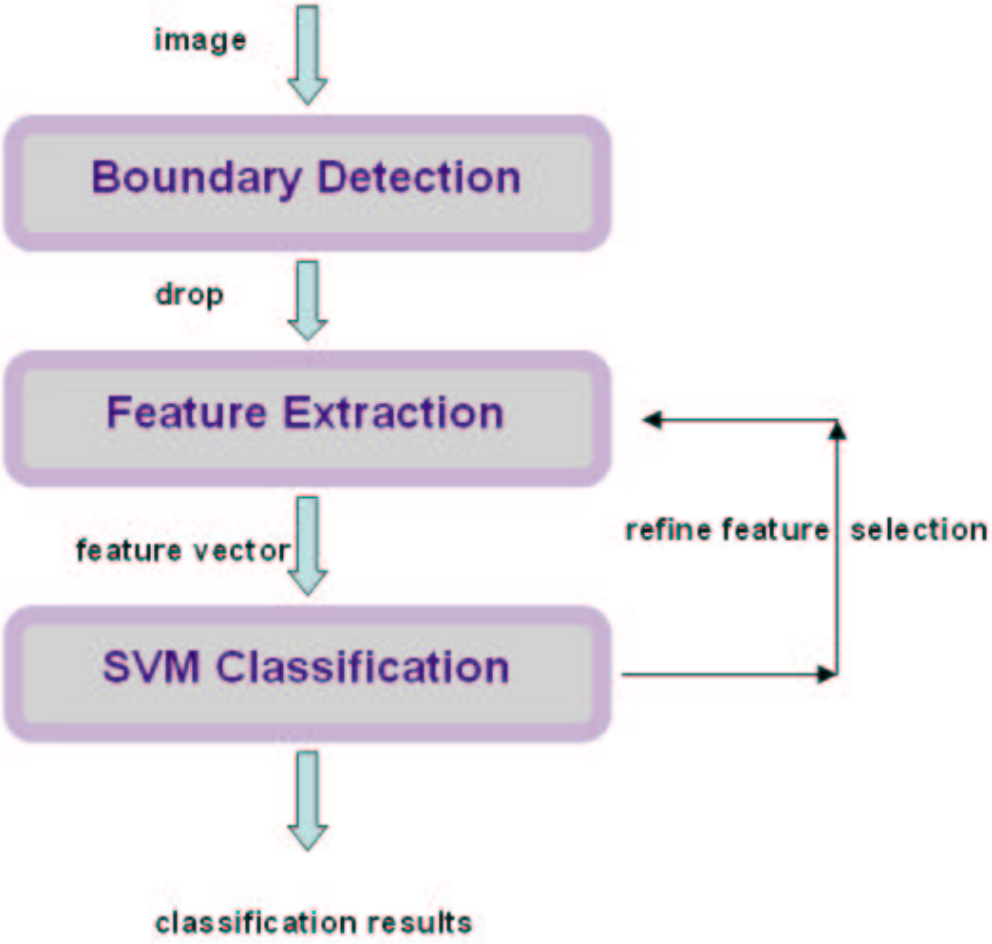

Figure 3: System diagram

3.2 Feature Extraction

As known to all, feature selection is usually the crucial step for a successful classification system. In practice, however, no universal rules exist. Features are usually selected in a heuristic manner, based on prior knowledge of the application and validated by numerous experiments. This is in fact the situation for classification of crystallization images. Features that have been tried in previous researches are mostly related to the geometric characteristic of a local region of the image (e.g., straight-line detection from Hough transform, conic curve fitting or line tracking).

In this project, we shall address the bottleneck of the current system, i.e., to improve differentiation between precipitates and crystal results. we plan to add new features which reflect the global informaiton of the images, especially in terms of texture and frequency characteristics.

The over-complete wavelet transform, in the form of a "steerable pyramid" [13] will be considered to capture phase coherence over a number of scales of the original images. The phase and amplitude information from the wavelet transform of the drop interior can serve as new feature vectors. More specifically, the shape-adaptive discrete wavelet transform (SA-DWT) [14] can be used to calculate the wavelet transform only within the drop boundary. A shift-invariant version is also available to alleviate the sensitivity of the original transform to the location of the droplet within the well[15].

Furthermore, it has been observed that the precipitate images tend to contain fluffy substance, whereas the crystal images tend to contain more crispy textures. Therefore, we will extract texture features from the outcome images, using standard techniques such as the co-occurence matrix [9] or Gabor filtering [10][16]. The co-occurence matrix computes a joint histogram of the gray level of one pixel and another pixel at a given distance and direction. A bank of different texture filters with various distances and directions can be used to describe the texture content of the images. We can summarize the matrices in terms of eigenvalues, condition number, and so on.

3.3 Classification

In this step, we use some representative feature vectors for crystals and non-crystals as a training set to design a classifier for the classification of new test data. Some learning algorithm takes this training set as input and produces a classifier as output. The resulting classifier then is used to tell whether a new feature vector corresponds to a crystal or not. We use a Support Vector Machine (SVM) as our classifier.

The main idea of a support vector machine is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized. It hinges on two mathematical operations: nonlinear mapping of an input vector into a high-dimensional feature space that is hidden from both the input and output; and construction of an optimal hyperplane for separating the features discovered in the previous step [17].

For our problem, the goal is to find a separation hyperplane between crystals and non-crystals from the given training dataset. The hyperplane should work well on the new

unknown test data. To construct such a hyperplane, SVM minimizes the structural risk, given as the probability of misclassifying previously unseen data. Thus, the SVM is theoretically supposed to be able to work well on new test data. During training, all the information in the training set is gradually packed into a small number of support vectors. Only these vectors are used to classify new data. In this way, we can know what features are the most effective in distinguishing between the two classes. These features can be explicitly obtained by further observations on the support vector shape [18][11][19]. The SVM exploits the information in the training set optimally. Compared with hand-crafted classification, it eliminates the process of defining appropriate discrimination criteria.

The selection of kernel directly affects the performance of the SVM. The kernels most commonly used include linear function kernel, radial basis function kernel, polynomial kernel, etc. We will try different kernels and find an optimal one for our problem. In [18], the author proposes a feedback framework that uses support vectors in order to obtain the distinctive feature components more explicitly. For our project since no similar work has been done before, we will adopt this method in our problem, to try to find which features are particularly distinctive, which would be very instructive to both refining existing features and exploring new features.

The MATLAB code we use in the project is provided by Dr. Göktürk and will be modified further for different kernels and feature formats.

4 Preliminary Results

In the preliminary experiments, we used a dataset consisting of 40 images from the website of Joint Center for Structural Genomics [20]. The first 10 images of them are classified as clear; the second 10 images of them are classified as precipitate; the third 10 images are classified as crystals, and the last 10 images are classified as mountable crystals.

An 11-dimensional feature vector (as discussed above) was obtained for each image. So we have altogether 40 vectors which were used as feature vectors by the SVM classification algorithm with linear functions as kernel. We used a 10-fold cross validation method to test the classifier. In the 10-fold cross validation, we divided the whole dataset into 10 disjoint sets and implemented 10 experiments. Each time, one of the ten sets was used as test set and the other as training set. Comparing the test results with the human classifications, we got all the true positives, false positives, true negatives and false negatives. We also implemented the existing decision-tree algorithm and compared it with our algorithm.

From table 1, we can see that the SVM algorithm achieved an accuracy of 55% with 9 false negatives, while the decision tree algorithm achieved an accuracy of 40% with 12 false negatives. Since we should try to avoid false negatives in the design of classifier, the SVM classifier shows more promising performance for this 40-image dataset.

We also analyzed what happens when the zero-crossing in SVM classifier is replaced by a level crossing. When the level is decreased, more true crystals will be detected, but at the cost of more false positives. The FROC curve, which shows the percentage of true crystals detections versus false positive detections, is given in Fig. 4.

In the experiments, we also tried the radial basis function kernel, but it showed inferior

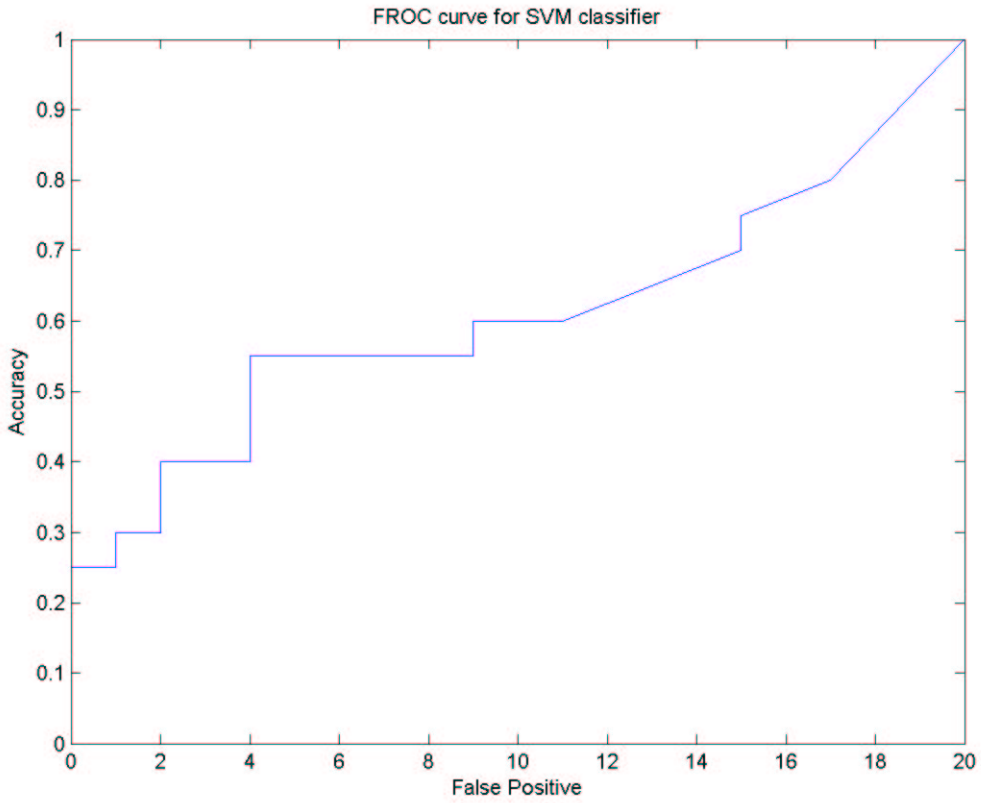

Figure 4: FROC curve

Table 1: Results on 40 images with the SVM versus the Decision Tree Classifier

	True Positive	False Negative	True Negative	False Positive
SVM	11	9	14	6
Decision Tree	8	12	15	5

performance compared to the linear kernel. We guess it may be because we have only 40 images as a training set and that may result in over fitting during training, which hurts the generality of the classifier.

5 Future Plan

Although the intial results show a promising trend in the use of the support vector machine as compared to the previous handcraft decision tree method, the classification accuracy achieved so far has yet to be improved. Note that no parameter tuning of the SVM has been performed yet, and that no new features are extracted at this point of time.

For future work, we plan to implement the level set method for boundary detection. The shape-adaptive wavelet transform will be applied to each crystallization image and several parameters will be calculated to serve as feature vectors. For instance, we can record the average energy level at different scales, the correlation factor of the phase information along different scales, as well as some statistical signatures such as the quantized distribution of high frequency coefficients. For the texture filters, we will calculate the co-occurence matrix for several different direction and distances, and record a summary vector for each matrix, e.g., the eigenvalues. Additionally, we will explore more about the RBF kernel in the SVM classifier and use the classification results to our feature selection.

6 Summary

Up till now, we have established an initial classification system for protein crystallization images based on some existing MATLAB and C code. We have substituted the original decision-tree classifier with the proposed SVM and have obtained preliminary results on a training set of 40 images. By this initial experiment, scripts for the cross-validation process are generated and will be useful for future tests with a larger dataset.

Future work will focus on the implementation of the level-set method, the wavelet and texture filters, as well as exploration of different kernels of SVM and feature selection from SVM classification feedback.

References

[1] I. Jurisica et al., "Intelligent decision support for protein crystal growth," *IBM Systems Journal*, 2001.

[2] W. M. Zuk and K. B. Ward, "Grant Proposal R21 and R33," *Journal of Crystal Growth*, 1991.

[3] J. Wilson, "Towards the automated evaluation of crystallization trials," *Acta Crystallographica D*, vol. 58, 2002.

[4] Marshall Bern et al., "Automatic classification of protein crystallization images using a line tracking algorithm," *Acta Crystallographica D*, 2003.

[5] Marshall Bern, "Grant Proposal R21 and R33," 2003.

[6] G. Spraggon, A. Kreusch S. A. Lesley, and J. P. Priestle, "Computational analysis of crystallization trials," *Acta Crystallographica D*, vol. 58, November 2002.

[7] G. Spraggon, A. Kreusch S. A. Lesley, and J. P. Priestle, "Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plate," November 2002.

[8] R. Malladi and J.A.Seithan, "Image processing: flows under min/max curvature and mean curvature," *Graphical Models and Image Processing*, vol. 58, 1996.

[9] R. Haralick, K. Shanmugam, and I. Dinstei, "Textural Features for Image Classification," *IEEE Transactions on Systems Man Cybernetics(SMC-3)*, 1973.

[10] O. Pichler, A. Teuner, and B.J. Hosticka, "A comparison of texture feature extraction using adaptive Gabor filter, pyramidal and tree structured wavelet transforms," *Pattern Recognition*, vol. 29, no. 5, 1996.

[11] B. Schölkopf, "Support vector learning," *R. Oldenbourg Verlag, Munich*, 1997.

[12] Weeratunga S. and C. Kamath, "An investigation of implicit active contours for scientific image segmentation," *Video Communications and Image Processing, SPIE Electronic Imaging*, January 2004.

[13] E.P.Simoncelli, W.T.Freeman, E.H.Adelson, and D.J.Heeger, "Shiftable multi-scale transforms," *IEEE Transaction on Information Theory, Special Issue on Wavelets*, vol. 38, 1992.

[14] S. Li and W. Li, "Shape-Adaptive Discrete Wavelet Transforms for Arbitrarily Shaped Visual Object Coding," *IEEE Transaction on Circuit and Systems for Video Technology*, vol. 10, pp. 725–743, August 2000.

[15] S. Del Marco, P. Heller, and J. Weiss, "An M-band 2-dimensional translation-invariant wavelet transform and applications," *Proc. of the 20th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-95), Detroit, Michigan*, pp. 1077–1080, May 1995.

[16] P. Kruizinga, N. Petkov, and S.E. Grigorescu, "Comparison of texture features based on Gabor filters," *Proceedings of the 10th International Conference on Image Analysis and Processing, Venice, Italy*, pp. 142–147, September 1999.

[17] Simon Haykin, "Neural networks: a comprehensive foundation," *Prentice Hall*, 1999.

[18] Salih Burak Göktürk and Carlo Tomasi, "A New 3-D Pattern Recognition Technique With Application to Computer Aided Colonoscopy," *Computer Vision and Pattern Recognition (CVPR'01)*, vol. 1, December 2001.

[19] V.N. Vapnik, "The nature of statistical learning theory," *Springer, New York*, 1995.

[20] "Joint Center for Structural Genomics," *http://www.jcsg.org.*