Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Integrated state evaluation for the images of crystallization droplets utilizing linear and nonlinear classifiers

Kuniaki Kawabata,[a]* Kanako Saitoh,[a,b] Mutsunori Takahashi,[b] Mitsuaki Sugahara,[a] Hajime Asama,[c] Taketoshi Mishima[b] and Masashi Miyano[a]

[a]RIKEN (The Institute of Physical and Chemical Research), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan, [b]Saitama University, 255 Shimo-okubo, Sakura, Saitama, Saitama 338-0825, Japan, and [c]The University of Tokyo, 5-1-5 Kashiwa, Kashiwanoha, Chiba 277-8568, Japan

Correspondence e-mail: kuniakik@riken.jp

Received 24 April 2006
Accepted 6 July 2006

In a usual crystallization process, the researchers evaluate the protein crystallization growth states based on visual impressions and repeatedly assign scores throughout the growth process. Although the development of crystallization robotic systems has generally realised the automation of the setup and storage of crystallization samples, evaluation of crystallization states has not yet been completely automated. The method presented here attempts to categorize individual crystallization droplet images into five classes using multiple classifiers. In particular, linear and nonlinear classifiers are utilized. The algorithm is comprised of pre-processing, feature extraction from images using texture analysis and a categorization process using linear discriminant analysis (LDA) and support vector machine (SVM). The performance of this method has been evaluated by comparing the results obtained using the method with the results obtained by a human expert and the concordance rate was 84.4%.

1. Introduction

For high-throughput protein structure determination by X-ray crystal structure analysis, the crystallization process is the first requisite. Therefore, there have been several attempts to achieve high-throughput crystallization (Stevens, 2000; Sugahara *et al.*, 2002). Through the development of robotic systems, the efficiency of crystallization work has seen rapid improvements. However, the observation process has not yet been completely automated, so that researchers must still keep samples under observation.

In previous work, methods related to automating the evaluation of crystallization droplets have been proposed (Bodenstaff *et al.*, 2002; Echalier *et al.*, 2004; Cumbaa *et al.*, 2003; Zuk & Ward, 1991; Rupp, 2003; Gester *et al.*, 2003). All these publications mainly discuss discrimination between the presence or absence of a crystal.

However, in addition to detecting crystals, it is also very important and valuable to observe and evaluate the crystallization growth states from start to finish. Several methods for evaluating crystallization growth states have been proposed. Spraggon *et al.* (2002) used texture analysis and a self-organizing neural network to categorize individual crystal trials into six classes. Bern *et al.* (2004) used the Hough transform and curve tracking and classified several states as 0 (empty), 1 (clear), 2 (precipitate), 3 (microcrystal) or 4 (crystal) using *C*5.0. Jurisica *et al.* (2001) used a two-dimensional Fourier transform and classified five distinct classes. Adams *et al.* (2002) extracted 11 features from images acquired by a RoboMicroscope II system and classified the

© 2006 International Union of Crystallography
Printed in Denmark – all rights reserved

droplets into four classes. Wilson (2002) utilized Bayes theorem and categorized the objects in the crystallization droplets into three classes based on size, shape, curvature of the boundary and the variance in intensity *etc.* Miyatake *et al.* (2005) also developed an automated crystallization/observation robotic system, HTS-80, which was reported to be able to categorize the crystallization droplet status into four stages based on extracted contour information.

In our previous work (Saitoh *et al.*, 2005), we reported the evaluation of incipient growth states of protein crystallization using texture information from greyscale crystallization images and statistical analysis, with the aim of achieving a categorization accuracy of more than 80% in each individual class. This method is one of the most highly efficient methods of evaluating the crystallization state.

However, these methods apply a single classifier to evaluate the image of the crystallization droplet and there was no consideration of the possibility of utilizing more than one classifier for a more accurate evaluation. Generally, in terms of statistical theory, an accurate performance can be achieved for a decision from discrimination results using multiple classifiers. Therefore, it is expected that a smaller evaluation error will be realised by applying multiple classifiers.

An integrated evaluation method that utilizes two classifiers is presented and the experimental results are reported. The method evaluates crystallization states to combine the computed results from both a linear classifier and a nonlinear classifier. LDA is used as the linear classifier and SVM is utilized as the nonlinear classifier.

2. Crystallization growth evaluation utilizing images

Structural genomics projects are expected to produce a large number of proteins from various organisms every year and so an efficient protocol will need to be established to obtain crystals suitable for X-ray structural analysis. Sugahara *et al.* (2002) have developed a fully automated protein crystallization and observation system, TERA. Fig. 1 shows examples of crystallization droplet images taken by TERA.

The growth states of protein crystallization take many aspects; for example, precipitate, amorphous agglutinate, crystals with varied shapes and combinations of these. In order to respond to these variations, RIKEN has developed ten standard categories for evaluation, as shown in Fig. 2.

A brief description of each RIKEN standard category is given below.

0, clear drop
1, precipitate (i). Creamy and grainless precipitate.
2, precipitate (ii). Fine or granulated sugar-like precipitate.
3, precipitate (iii). Amorphous state.
4, amorphous grain. Circular grain
5, microcrystal (size of 50 μm or less).
6, crystal (i). Needle crystal or plate crystal.
7, crystal (ii). Cluster of crystals.
8, crystal (iii). Single crystal (0.05–0.2 mm).
9, crystal (iv). Single crystal (greater than 0.2 mm).

Our target classification, which is a reclassification of the RIKEN system, is shown in Fig. 2 and is the same as that used for categorization in our previous work (Saitoh *et al.*, 2005). Categories 4 (amorphous grain) to 9 [crystal (iv)] are placed together into one category (*E*) and all samples are classified into five categories from *A* to *E*.

3. Pre-processing and feature extraction

The method used for evaluation of the crystallization growth states uses a process flow consisting of feature extraction and classification. Generally, in the pre-processing phase the input images are processed for noise, normalization and so on. Feature extraction entails the extraction of features used for classification from the original images. The classification process classifies the input images using these features.

Before the feature-extraction process, some pre-process steps are performed. The original images are photographed by TERA using a microscope at 40× magnification. The image size is 1392 × 1040 pixels. Initially, the original colour images are transformed into 256-level greyscale images because the colour information is not utilized in the method. A portion of the original image from inside the well is then manually extracted. The object used for processing in this study is assumed to only be inside the well. The extract size is 150 × 150 pixels, which is determined by considering an approximated average size for microcrystals and crystals in the original images. Finally, the extracted image is differentiated with a Sobel first-order differential filter. This process highlights the characteristic pattern of the image. Both differentiated and non-differentiated images are utilized in this method.

In the computational image analysis, texture analysis, smoothing, sharpening and edge-detection/enhancement are often utilized for quantifying images (Takagi & Shimoda, 1991). In our proposed method (Saitoh *et al.*, 2005), we introduced texture analysis (Haralick *et al.*, 1973), which quantifies the array of greyscale values of each pixel in an image for use in extracting features from the crystallization images. In our previous work (Saitoh *et al.*, 2005), we have already discussed those features that are effective for accurate classification.

Texture-feature values are calculated using a grey-level co-occurrence matrix. Fig. 3 shows the algorithm for deriving the matrix. Each element of the matrix $P_\delta(i, j)$ in Fig. 3(*b*) expresses the probability that the greyscale value of a pixel is i and the greyscale value of another pixel located at r pixels away in the θ direction from the former pixel is j (Fig. 3*a*). The displacement between the two pixels is denoted $\delta = (r, \theta)$. The distance between pixels, r, used to calculate the co-occurrence matrices has a value of 1. The directions θ are 0, 45, 90, 135° and correspond to the horizontal, the vertical and two diagonals. If nothing is done, the features are calculated anisotropically. However, the crystallization growth states are not anisotropic, so the average of the results calculated from the four directions is used. 14 texture features can be calculated using the matrix P_δ.

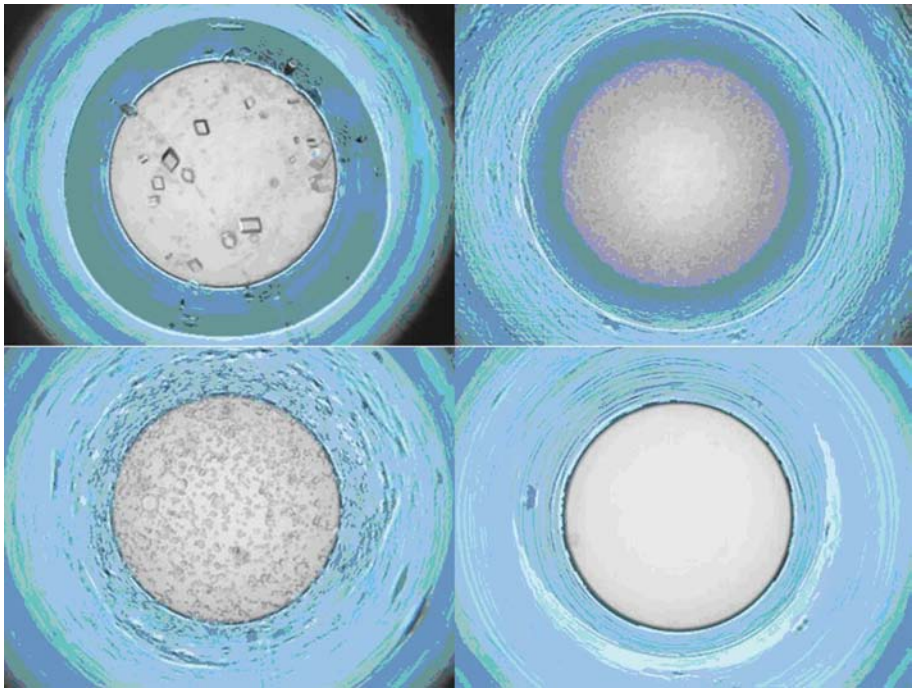

Figure 1
Examples of drop images taken with TERA (developed by RIKEN). The image size is 1392 × 1040 pixels and the well is visible in the centre.

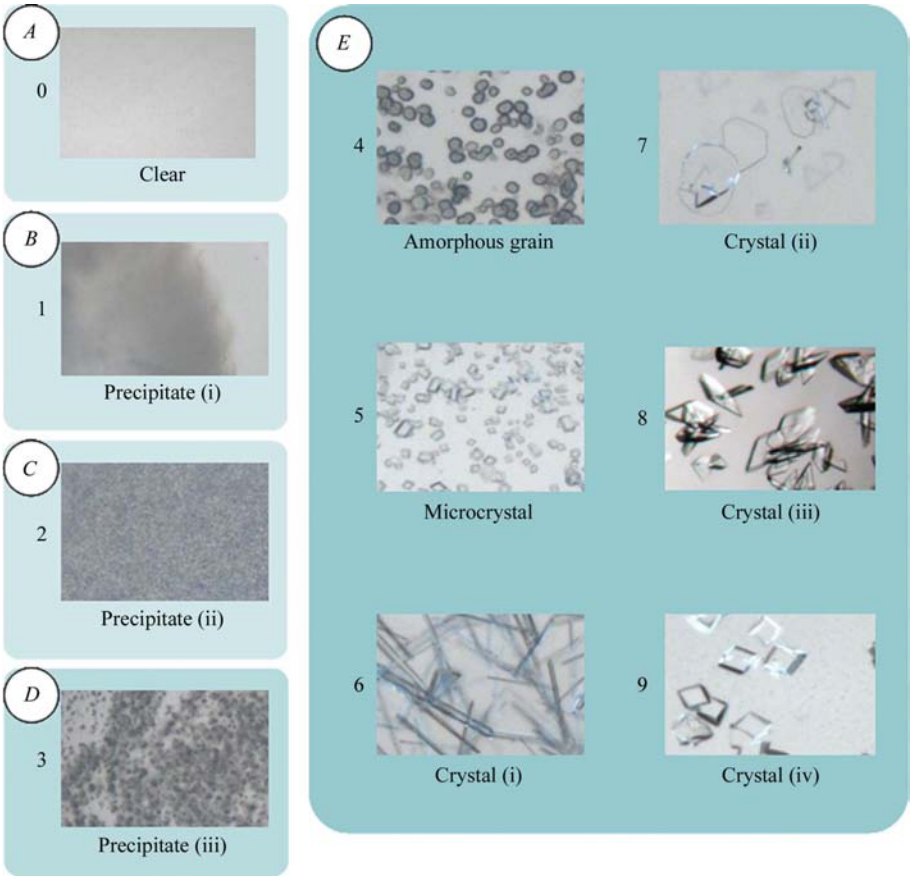

Figure 2
The ten categories for evaluation set by RIKEN incorporated by the target categories. *A*, *B*, *C*, *D* and *E* indicate the target categorization in this paper. Images are taken from a part of the well in the full drop images.

As is mentioned above, in this process both differentiated and non-differentiated images are utilized for classification. Thus, the total number of types of feature values is 28 (14 are extracted from the differentiated image and 14 from the non-differentiated image).

4. Classifiers and classification procedure

It is possible to use any classifier as the element classifier of each node. In this section, the linear and nonlinear classifiers which are utilized in the method are presented and the basic classification algorithm is described.

4.1. Classifiers

4.1.1. Linear discriminant analysis (LDA). In this method, LDA, which is a standard technique for multivariate analysis, is used as the linear classifier to classify the feature vector.

LDA (Fisher, 1936) is used to compute a linear discriminant function that divides the feature space into two groups. A discriminant space is constructed from the linear transformation

$$^{\mathrm{LDA}}g(\mathbf{u}) = \mathbf{A}^T\mathbf{u} + a_0,$$

where $\mathbf{A}$ is a coefficient matrix, $\mathbf{u}$ is a texture-feature vector and a_0 is a constant involving $\mathbf{A}$. The coefficient matrix $\mathbf{A}$ is computed so that the discriminant criterion

$$J_\Sigma(\mathbf{A}) = \frac{\mathbf{A}^T\Sigma_B\mathbf{A}}{\mathbf{A}^T\Sigma_W\mathbf{A}},$$

may be maximized, where Σ_B and Σ_W are the between-class covariance matrix and within-class covariance matrix, respectively. To classify a new input vector, $g(\mathbf{u})$ is computed to determine the class, depending on its positive or negative sign (Fig. 4).

4.1.2. Support vector machine (SVM). SVM (Vapnik, 1995) is applied as the nonlinear classifier. SVM is a technique that is well founded in statistical learning theory and is used to train classifiers, regressors and probability densities (Fig. 5). One of the main attractions for using SVM is that it

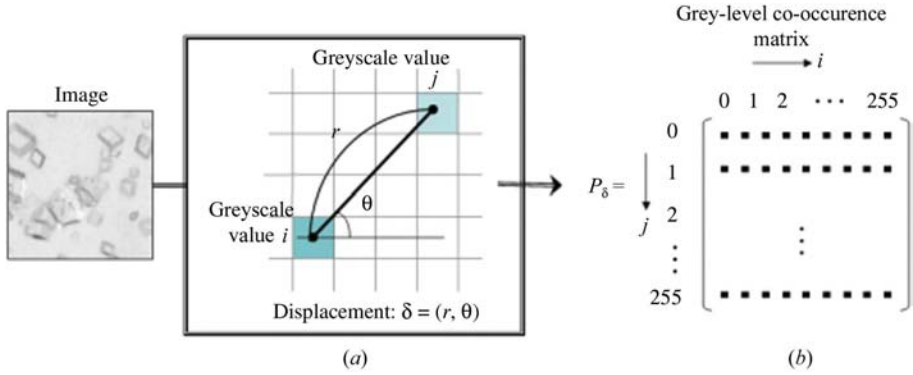

Figure 3
Algorithm used to derive the grey-level co-occurrence matrix. Texture-feature values are then calculated using the grey-level co-occurrence matrix [P_δ in (*b*)]. (*a*) Each element of the matrix expresses the probability that the greyscale value of one pixel is *i* and the greyscale value of another pixel located *r* pixels away in the θ direction from the former pixel is *j*. The parameters $r = 1$ and $\theta = 0$, 45, 90 and 135° are used in this work.

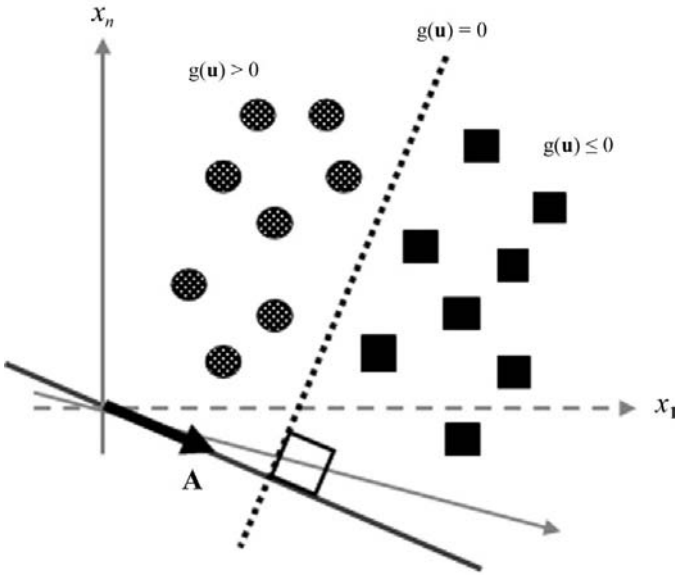

Figure 4
Linear discriminant analysis. $g(\mathbf{u})$ divides the feature space into two groups. A discriminant space is constructed from the linear transformation $g(\mathbf{u}) = \mathbf{A}^T\mathbf{u} + a_0$. Input data are classified depending on their sign.

is capable of learning in sparse high-dimensional spaces with very few training examples. SVM accomplishes this by simultaneously minimizing the bounds of empirical error and the complexity of the classifier. The following is a brief overview of the main concepts of SVM.

SVM performs pattern recognition for two-class problems by determining the separating hyperplane with a maximum distance to the closest points of the training set. These points are called support vectors. If the data is not linearly separable in the input space, a non-linear transformation $\Phi(\cdot)$ can be applied which maps the data points $\mathbf{u}$ into a high-dimensional space H, which is called the feature space. The data in the feature space is then separated by the optimal hyperplane described above.

The mapping transformation, $\Phi(\cdot)$, is represented in the SVM classifier by a kernel function, $K(\cdot, \cdot)$, which defines an inner product in H; *i.e.* $K(\mathbf{u}, \mathbf{u}') = \Phi(\mathbf{u})^T\Phi(\mathbf{u}')$. The determining function of the SVM has the form

$$^{\mathrm{SVM}}g[\Phi(\mathbf{u})] = \sum_{i=1}^{m} a_i y_i K(\mathbf{u}, \mathbf{u}_i),$$

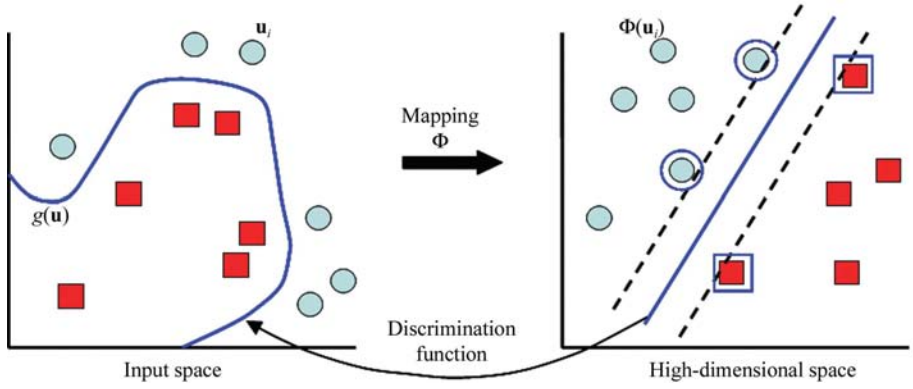

Figure 5
Support vector machine (SVM) is a technique that is well founded in statistical leaning theory and is used to train classifiers, regressors and probability densities.

where m is the number of data points and $y_i \in \{-1, 1\}$ is the class label of training points $\mathbf{x}_i$. The coefficient α_i can be found by solving a quadratic with linear constraints. The support vectors are the nearest points to the separating boundary and are the only ones for which α_i can be non-zero.

Examples of admissible kernel functions are the polynomial kernels

$$K(\mathbf{u}, \mathbf{u}') = (\mathbf{u}^T\mathbf{u}' + 1)^d,$$

where d is the degree of the polynomial, and the Gaussian kernels

$$K(\mathbf{u}, \mathbf{u}') = \exp(-||\mathbf{u} - \mathbf{u}'||/2\sigma^2),$$

where σ is the variance of the Gaussian function. In this work, the Gaussian kernel is utilized.

4.2. Classification procedure and classification by each classifier

Our target categorization is five classes from *A* to *E*. In order to realise this categorization, the classifier is applied step by step for two classes. Four discriminant functions, ${}^jg_1$, ${}^jg_2$, ${}^jg_3$, ${}^jg_4$ (j = LDA, SVM), are computed by each classifier. The grouping and the sequence of the discrimination process were decided as shown in Fig. 6. When the functions ${}^jg_1$, ${}^jg_2$ and ${}^jg_4$ (j = LDA, SVM) are derived, the texture-feature values calculated from differentiated images are used. In case of the function ${}^jg_3$ (j = LDA, SVM), the feature values are calculated from non-differentiated images.

For classification, both of differentiated and non-differentiated images are utilized. 14 feature values are derived from each type of image. The total number of kind of feature values is 28.

The classification experiments are performed using the abovementioned classifiers. The performance of each classifier is evaluated based on how the classification results are in concordance with those provided by a human expert.

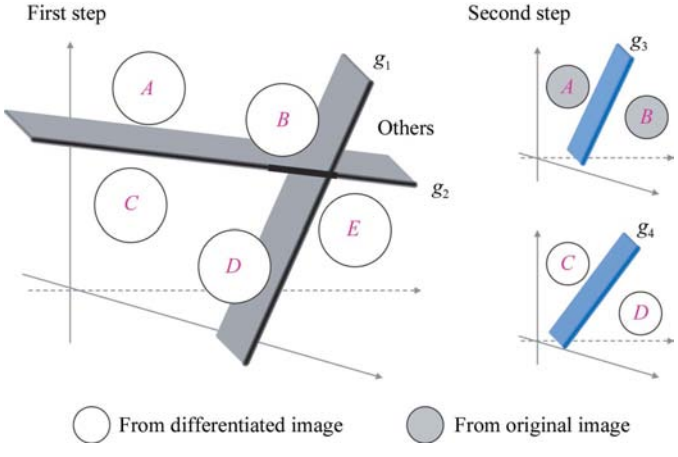

Figure 6
Discrimination procedure. The procedure consists of two steps. (1) The input data is classified into *A*, *B* or *C*, *D* or *E* using the functions ${}^jg_1$ and ${}^jg_2$. (2) The data is classified into *A* or *B* or *C* or *D* using the functions ${}^jg_3$ and ${}^jg_4$ (j = LDA, SVM)

Table 1
Results of classification with LDA.

The results obtained using LDA are compared with the results obtained by an expert (manual classification). The overall accuracy is 80.47%.

			Classification with LDA					
		Total	*A*	*B*	*C*	*D*	*E*	*F*
Manual classification	*A*	53	**51**	1	1	0	0	0
	B	71	3	**52**	15	0	1	0
	C	29	0	1	**22**	4	2	0
	D	41	0	0	2	**31**	8	0
	E	245	0	1	2	13	**229**	0
Accuracy (%)			94.44	94.55	52.38	64.58	95.42	

Table 2
Results of classification with SVM.

The results obtained using SVM are compared with the results obtained by an expert (manual classification). The overall accuracy is 80.07%.

			Classification with SVM					
		Total	*A*	*B*	*C*	*D*	*E*	*F*
Manual classification	*A*	53	**51**	2	0	0	0	0
	B	71	0	**46**	21	2	2	0
	C	29	0	0	**20**	2	7	0
	D	41	0	0	0	**27**	14	0
	E	245	0	0	1	10	**234**	0
Accuracy (%)			100.0	95.83	47.62	65.85	91.05	

Table 3
Results of classification using both LDA and SVM.

The results obtained using both LDA and SVM are compared with the results from an expert (manual classification). The overall accuracy is 84.84%.

			Classification with LDA and SVM					
		Total	*A*	*B*	*C*	*D*	*E*	*F*
Manual classification	*A*	53	**50**	1	0		0	2
	B	71	0	**44**	12	0	1	14
	C	29	0	0	**17**	1	1	10
	D	41	0	0	0	**25**	7	9
	E	245	0	0	1	8	**229**	7
Accuracy (%)			100.0	97.78	56.67	73.53	96.22	

The data set for performance evaluation contains 874 images obtained with TERA that were annotated by a human expert at RIKEN. The number of images in each category is as follows: *A* (clear), 102; *B* [precipitate (i)], 116; *C* [precipitate (ii)], 78; *D* [precipitate (iii)], 90; E (amorphous grain, microcrystal, crystal), 488. Category *E* has over five times more images than the other categories, because the images were acquired to maintain an approximately constant number from each of the ten categorizations set by RIKEN. Of the total 874 images, 435 (*A*, 49; *B*, 45; *C*, 49; *D*, 49; *E*, 243) images were used as the training set and 439 (*A*, 53; *B*, 71; *C*, 29; *D*, 41; *E*, 245) were used as the test set. Those training images in which some artifacts were present in the crystallization drop were removed in advance.

The evaluation results using LDA with the test set are given in Table 1. The results obtained using LDA were compared with the results from the human expert (manual classification; Table 1) and the concordance rates were calculated by the

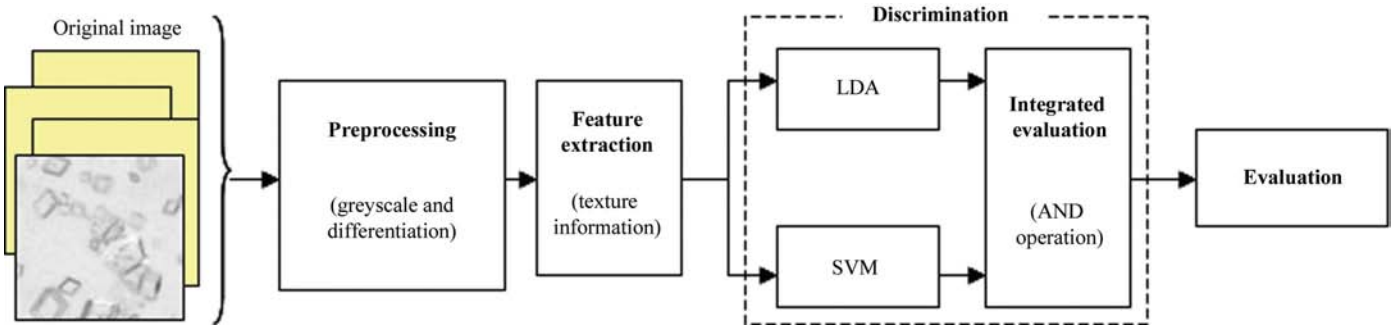

Figure 7
Integrated evaluation system. Extracted feature values (texture information) are utilized for evaluation by LDA and SVM. Final evaluation is derived by the logical AND operation between the results of LDA and SVM.

expression (accuracy) = (the number of images classified correctly)/(total number) × 100. Class *F* (Table 1) was designated for samples that were not categorized into any other classes (*A*–*E*). Of all the images, 80.47% were classified into the same category as manually classified by an expert.

The evaluation results obtained using SVM with the test set are given in Table 2. The parameters used for the SVM are as follows: for ${}^{\rm SVM}g_1$, $\sigma = 0.01$, $c = 200$; for ${}^{\rm SVM}g_2$, $\sigma = 0.01$, $c = 200$; for ${}^{\rm SVM}g_3$, $\sigma = 0.1$, $c = 20$; for ${}^{\rm SVM}g_4$, $\sigma = 0.1$, $c = 1000$, where σ is the variance of the Gaussian function and c indicates the SVM misclassification tolerance parameter.

The parameters were determined by estimating the performance of each function. The results obtained using SVM were compared with the results obtained from the human expert. The average accuracy of classification obtained using the SVM is 80.07%.

5. Integrated evaluation combining the results from two classifiers

In this section, an integrated evaluation method is discussed. The linear and non-linear discrimination methods have been described. Finally, these two classifiers are combined to achieve a more accurate evaluation.

The proposed method employs LDA and SVM, which are both noteworthy for their efficiency in two-class identification. Fig. 7 shows the evaluation process of the method. The evaluation sequence is as follows.

(i) Pre-process to cutoff image and generate greyscale and differentiated images.

(ii) Extract texture information (14 feature values) from pre-processed image (both of differentiated image and non-differentiated image). This is performed to extract 28 feature values from each cutoff image.

(iii) Evaluate each image utilizing the extracted feature values from both of the classifiers (derived by LDA and SVM algorithms), independently.

(iv) Compare the results from both of the classifiers for integrated evaluation. When the results of LDA and SVM processing are equivalent, the evaluation is adopted. When the results are not equivalent, the evaluation is denoted class *F* (unknown).

The evaluation result obtained using the test set is given in Table 3. The results obtained using the combined LDA and SVM classification method were compared with the results from a human expert. Of all the images, 84.84% were classified into the same category as that manually classified by an expert.

6. Summary

A method for the evaluation of protein crystallization states based on two classifiers has been presented. The method classifies the images into five groups. The images taken by the automated crystallization system TERA are used and a texture-analysis method is utilized to extract feature values from each image.

In our previous work, a basic classification procedure was presented based on texture information and linear discriminant analysis that utilized step-by-step classification of samples into five groups. In this work, the method is extended to utilize two classifiers for a more accurate classification. Evaluation results from each classifier are compared with each other and when the score is the same, the score is accepted as an acceptable evaluation.

Of those images classified (100% successful classification without unknowns), the method provides 84.84% accuracy and concordance with expert observations. The results from combined classification show a more accurate collection performance than the results using each classifier individually.

Future work will consider examination of how to combine plural classifiers to realise more accurate evaluation.

This work has been partially supported by a research grant of the Okawa Foundation for Information and Telecommunication (05-23).

References

Adams, J. A., Jewell, D., Jorgensen, K., Mickly, M. & Newman, J. M. (2002). *JALA*, **7**, 36–40.
Bern, M., Goldberg, D., Stevens, R. C. & Kuhn, P. (2004). *J. Appl. Cryst.* **37**, 279–287.
Bodenstaff, E. R., Hoedemaeker, F. J., Kuil, M. E., de Vrind, H. P. M. & Abrahams, J. P. (2002). *Acta Cryst.* D**58**, 1901–1906.

Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., DeTitta, G. & Jurisica, I. (2003). *Acta Cryst.* D**59**, 1619–1627.
Echalier, A., Glazer, R. L., Fülöp, V. & Geday, M. A. (2004). *Acta Cryst.* D**60**, 696–702.
Fisher, R. A. (1936). *Ann. Eugen.* **7**, 179–188.
Gester, T. E., Rosenblum, W. M., Christopher, G. K., Hamrick, D. T., DeLucas, L. J. & Tillotson, B. (2003). United States Patent 6 529 612.
Haralick, R. M., Shanmugam, K. & Dinstein, I. (1973). *IEEE Trans. Syst. Man. Cybern.* **SMC-3**, 610–621.
Jurisica, I., Wolfley, J. R., Rogers, P., Bianca, M. A., Glasgow, J. I., Weeks, D. R., Fortier, S., DeTitta, G. & Luft, J. R. (2001). *IBM Syst. J.* **40**, 394–409.
Miyatake, H., Kim, S.-H., Motegi, I., Matsuzaki, H., Kitahara, H., Higuchi, A. & Miki, K. (2005). *Acta Cryst.* D**61**, 658–663.
Rupp, B. (2003). *Acc. Chem. Res.* **36**, 173–181.
Saitoh, K., Kawabata, K., Asama, H., Mishima, T., Sugahara, M. & Miyano, M. (2005). *Acta Cryst.* D**61**, 873–880.
Spraggon, G., Lesley, S. A., Kreusch, A. & Priestle, J. P. (2002). *Acta Cryst.* D**58**, 1915–1923.
Stevens, R. C. (2000). *Curr. Opin. Struct. Biol.* **10**, 558–563.
Sugahara, M., Nishio, K., Kobayashi, M., Hamada, K. & Miyano, M. (2002). *Internation Conference on Structural Genomics, Berlin, Germany*. Abstract ID 69.
Takagi, M. & Shimoda, H. (1991). *Handbook Of Image Analysis.* Tokyo: University of Tokyo Press.
Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer–Verlag.
Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.
Zuk, W. M. & Ward, K. B. (1991). *J. Cryst. Growth*, **110**, 148–155.