Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates

Christian A. Cumbaa,[a] Angela Lauricella,[b] Nancy Fehrman,[b] Christina Veatch,[b] Robert Collins,[b] Joe Luft,[b] George DeTitta[b] and Igor Jurisica[a]*

[a]Ontario Cancer Institute, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada, and [b]Hauptman–Woodward Medical Research Institute, 73 High Street, Buffalo, NY 14203-1196, USA

Correspondence e-mail: ij@uhnres.utoronto.ca

A technique for automatically evaluating microbatch (400 nl) protein-crystallization trials is described. This method addresses analysis problems introduced at the sub-microlitre scale, including non-uniform lighting and irregular droplet boundaries. The droplet is segmented from the well using a loopy probabilistic graphical model with a two-layered grid topology. A vector of 23 features is extracted from the droplet image using the Radon transform for straight-edge features and a bank of correlation filters for microcrystalline features. Image classification is achieved by linear discriminant analysis of its feature vector. The results of the automatic method are compared with those of a human expert on 32 1536-well plates. Using the human-labeled images as ground truth, this method classifies images with 85% accuracy and a ROC score of 0.84. This result compares well with the experimental repeatability rate, assessed at 87%. Images falsely classified as crystal-positive variously contain speckled precipitate resembling microcrystals, skin effects or genuine crystals falsely labeled by the human expert. Many images falsely classified as crystal-negative variously contain very fine crystal features or dendrites lacking straight edges. Characterization of these misclassifications suggests directions for improving the method.

1. Introduction

High-throughput robotic protein-crystallization systems are now capable of testing thousands of protein-cocktail combinations per day. Problems and challenges with 1536-well systems have been discussed in general, *e.g.* Garyantes (2002). In this paper, however, we focus on the problem of automatically evaluating the outcome of each trial, a problem relevant even for lower-throughput systems. This problem is a rate-limiting step in the search for favourable protein-crystallization conditions, ordinarily requiring a human expert (crystallographer) to visually assess the outcome of each trial. Here, we describe progress in automating this evaluation step.

The choice of tool in a search for crystallization conditions should reflect the task. A crystallographer who uses a 1536-condition screen casts a wide net across the crystallization-condition space. Often, the proteins in these screens do not crystallize easily and any crystal outcome, even microcrystalline precipitate, is a positive result. Thus, we designed our method to detect these borderline outcomes as well as larger well formed crystals.

Several recent publications discuss automatic classification of images from 96-well (Spraggon *et al.*, 2002; Wilson, 2002; Rupp, 2003; Adams *et al.*, 2002) and 1536-well (Jurisica, Rogers, Glasgow, Collins *et al.*, 2001) crystallization screens.

Our work extends the limits of feasible automated high-throughput protein crystallization screening to the sub-microlitre drop level. Bodenstaff *et al.* (2002) discuss an alternative method for crystal detection in sub-microlitre trials which avoids significant image analysis by using rotating polarizing filters to highlight birefringent crystals. As in Spraggon *et al.* (2002) and Adams *et al.* (2002), we classify the state of the entire droplet; in contrast, the method of Wilson (2002) classifies individual objects within the drop. As in Rupp (2003) and Bodenstaff *et al.* (2002), we restrict our evaluation to a binary set of outcomes (crystals or no crystals). Comparing image-analysis methods, the methods of Spraggon *et al.* (2002) and Wilson (2002) use the Hough transform to detect straight-line features (crystals) in their images. We use the Radon transform, which is similar to the Hough transform, for the same purpose. Unique to our method, we also employ a correlation filter to detect microcrystals. These image-analysis methods are all described in §2.3.

Our data comes from the robotic pipetting and imaging system at the Hauptman–Woodward Medical Research Institute (HWI; Luft *et al.*, 2001). The robotic crystallization system at HWI has changed since the description of an earlier image-analysis method developed for the system (Jurisica, Rogers, Glasgow, Collins *et al.*, 2001). The drop volume per well in the 1536-well plates has decreased from 1.0 µl to 400 nl. The camera position has also changed: plates are now photographed from below and thus the drops are now seen through the floor of the plate instead of through the layer of mineral oil that caps each well.

These changes have introduced factors that complicate image analysis. Two changes are especially significant: reversal of the camera position and reduction of the droplet volumes.

Photographing experiments from the bottom of the well makes crystal identification by crystallographers easier. It improves the quality of the image, eliminating the image distortion that occurs when viewing the experiment drop through a layer of oil. However, reversing the camera position introduces the back-lit curved surface of the oil barrier in the well which acts as a lens, concentrating light in the centre of the well and leaving the edges dark. As a result, objects in the centre of the well are relatively dark and contrast sharply with the light background; objects to the sides of the well are relatively light but contrast poorly with the dark background. Photographing the droplets from below also filters the light through the non-optical-grade plastic floor of the well, introducing rippled bands of alternating light and dark into the image, as shown in Fig. 1.

Decreasing the droplet volume (while using the same well) changes the drop contour. Whereas droplets were approximately elliptical in shape before, they now frequently have irregular non-convex contours as depicted in Fig. 1. Combined with the non-uniform lighting, droplet segmentation is now highly non-trivial and rivals the difficulty of the classification problem. [In contrast, Spraggon *et al.* (2002), Wilson (2002) and Jurisica, Rogers, Glasgow, Collins *et al.* (2001) each faced the simpler problem of approximately circular drops against a near-uniform background; each solved it with a simple geometric technique.] We describe a probabilistic model of the droplet boundary in §2.2.

2. Methods

Automatic image classification comprises several stages: *well registration* locates and eliminates the boundaries of the well, *droplet segmentation* eliminates the edges of the drop and *feature extraction* extracts descriptive features used during *image classification*.

2.1. Well registration

The first step in the image-analysis problem is to locate the boundaries of the well so that subsequent analysis may be performed on its contents. Each well in the 1536-well plates has the same dimensions and is photographed individually by HWI's robotic imaging system. The result is a greyscale 632×504 image, with the well occupying a 425×425 pixel square somewhere in the center. The edges of the well are straight and are generally aligned within a degree or two of the image axes.

We incorporate a previously described registration method (Jurisica, Rogers, Glasgow, Collins *et al.*, 2001). To obtain the vertical well boundaries, we find the pair of pixel columns separated by the expected well width with the closest average pixel intensity. We apply the same search through pixel rows (separated by the expected well height) to obtain the horizontal well boundaries.

2.2. Droplet segmentation

Before extracting features from the region of the image containing the droplet, we have to identify the edges of the droplet. This is the segmentation step. We use droplet segmentation to divide the well interior into three pieces: the
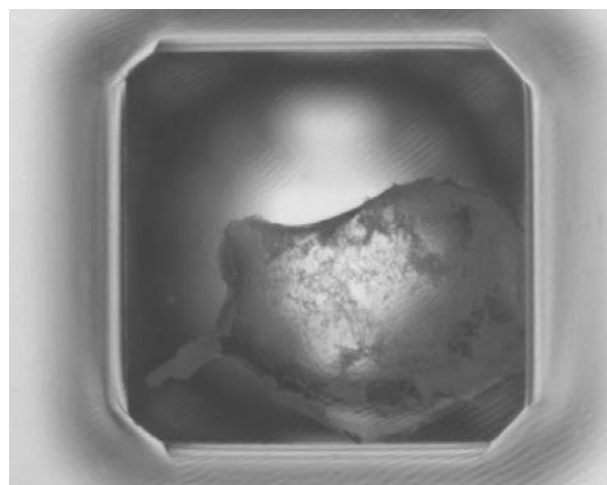

Figure 1
A sample image illustrating complicating factors in image analysis. Back-lighting causes non-uniform lighting (a bright centre surrounded by darkness) and reveals ripples on the surface of the well bottom (introducing spurious edges in the image). Small droplet volumes cause droplets to form irregular non-convex contours.

empty well (W), the inside of a droplet (D) and the edge of a droplet (E). This allows us to exclude the empty regions of the well and the edges of the droplet from further analysis. We include in the edge segment of an image any shadows or reflective or highly refractive portions on the droplet-well boundary. In most images, these regions are where the highest intensity gradients and most extreme intensity values occur. Thus, it is especially important that we exclude the droplet edges from the image-smoothness and straight-edge analyses described in §2.3. We do this at the risk of unintentionally discarding crystals hiding in or near the droplet edge.

We used a probabilistic graphical model to segment the central region of the well, illustrated in Fig. 2. We first divide the well into a grid of 17×17 regions. Each region (i, j) in this grid is represented by a pair of variables in the graphical model, forming a 60-component mixture of multivariate Gaussian distributions: one latent discrete variable $x_{i,j} \in (1, \ldots, 60)$ controlling the active mixture component and one observed vector-valued variable $\mathbf{y}_{i,j}$ representing local image intensity and gradient values computed directly from the image. Each mixture models the local state of the well, be it W, D or E. The state of region (i, j) is encoded in the value of the latent variable $x_{i,j}$, with 20 mixture components assigned to each possibility. (Each possible state is given multiple mixture components in order to cover the variety of textures possible in an underlying image region with a given state.) Each latent variable $x_{i,j}$ is also linked to its neighbors by conditional probability relationships $P(x_{i,j}|x_{i-1,j})$, $P(x_{i+1,j}|x_{i,j})$, $P(x_{i,j}|x_{i,j-1})$ and $P(x_{i,j+1}|x_{i,j})$. These probabilities, together with the conditional probability $P(\mathbf{y}_{i,j}|x_{i,j})$ of the mixture, ensure that the inferred value of $x_{i,j}$ depends on the local region of the image as well as the values of its neighbors.

We refer to the complete set of latent variables $\mathbf{X} = \{x_{i,j}\}$ as the latent layer of the model and the complete set of observed variables $\mathbf{Y} = \{\mathbf{y}_{i,j}\}$ as the observed layer of the model. A configuration $\mathbf{X} = \mathbf{x}$ of the latent-variable layer represents a possible segmentation of the well. Segmentation of a particular image is accomplished by inferring the most likely configuration from the probability distribution $P(\mathbf{X}|\mathbf{Y} = \mathbf{y})$, where $\mathbf{y}$ is computed directly from the image.

Inference in graphical models is efficient and exact when the graph has a tree topology. The model used in this study necessarily has a highly connected grid topology and so must use an inexact Loopy Belief Propagation algorithm (Murphy *et al.*, 1999) to infer the most likely segmentation. The message-passing schedule we employed amounts to several iterations of alternately and iteratively updating each row of $\mathbf{X}$ and then each column using the forward–backward algorithm.

We trained our segmentation model on 45 hand-segmented images. To reduce the number of parameters in the model, we tied all horizontal conditional probabilities $P(x_{i,j}|x_{i-1,j})$ to a single distribution and did the same for all vertical conditional probabilities $P(x_{i,j}|x_{i,j-1})$. We also tied all mixture model parameters together so that we only needed to train a single 60-component mixture model [*i.e.* a single mixture prior $P(x)$ and 60-component distributions of the observed variable $P(\mathbf{y}|x)$].

Validation of the image-segmentation model was performed on a set of 50 hand-segmented images containing 4319 empty-well regions, 1348 droplet-border regions and 8783 intra-droplet regions. The segmenter correctly identified 96% of all well regions, 69% of all border regions and 95% of all intra-droplet regions for a weighted mean of 93% overall accuracy.
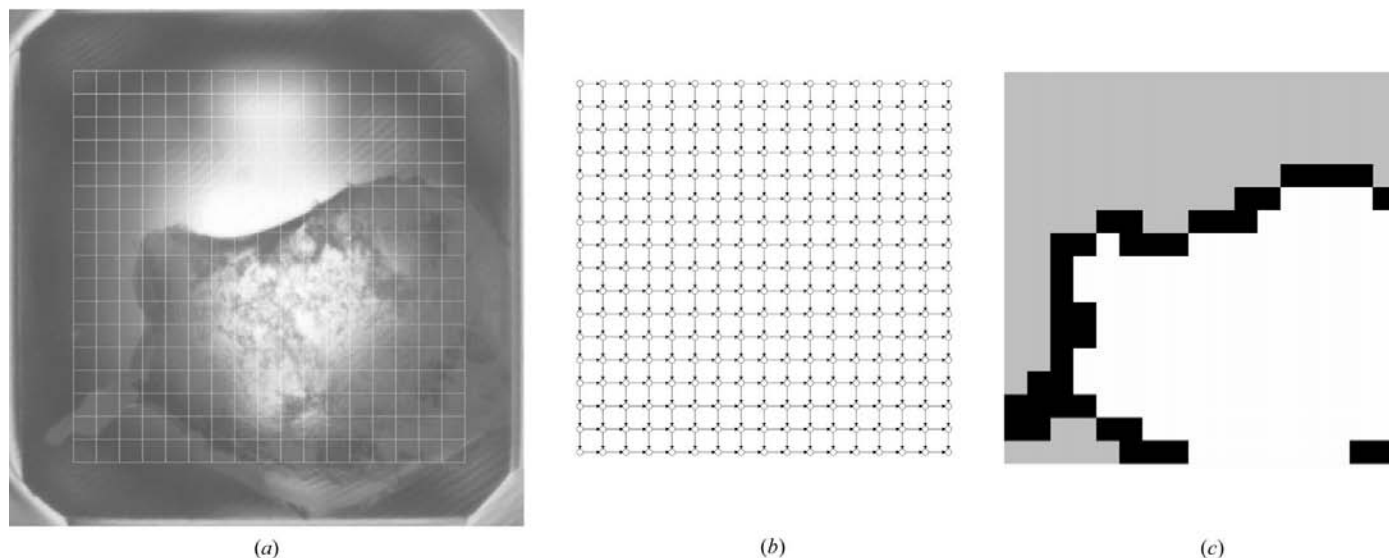

(a)   (b)   (c)

Figure 2
The Bayes net segmentation model. (*a*) We divide the well into a 17×17 grid of tiles and compute a vector of image-intensity gradients for each tile. These vectors are the observed values of variables $\mathbf{y}_{i,j}$. We model each tile's vector as a sample drawn from a mixture of multivariate Gaussians. (*b*) A mesh of conditionally interdependent latent variables, one per tile, control which component is active in each mixture. (*c*) The active component in each tile determines the segment (one of $\{W, D, E\}$) assigned to that patch of the image.

2.3. Feature extraction

A single image contains a lot of information and many different descriptive features can be extracted. However, much of the information is redundant or irrelevant to the classification task. The goal is to extract a minimal set of features that enables the classifier to discriminate among different crystallization experiment outcomes with high accuracy.

Protein crystals come in a great variety of forms: microcrystals, microneedles, needle crystals and larger faceted crystals. Although the case of microcrystals the image magnification may not be sufficient to discern these straight edges, most crystalline material will exhibit this feature. Straight edges in a droplet are therefore important features to detect. Owing to their small size, microcrystals have no straight edges of significant length (tens of pixels). We thus devised a set of image features specifically for detecting microcrystals. Similarly, there are a variety of precipitates, but the classification of them is beyond the scope of this paper.

In total, we reduce each image to a vector of 23 features: 20 measuring microcrystal features, two measuring the presence of straight edges detected within a droplet and one measuring the smoothness of the contents of the droplet.

2.3.1. Image-analysis tools. Two image-analysis tools form the computational basis of the image features computed by our method: the Laplacian operator and the Radon transform. The Laplacian operator computes the rate of change of the intensity gradient in an image, *i.e.* the second derivative of the intensity of an image. It is approximated by the following convolution that computes the summed differences between each pixel's intensity and its neighbors' intensities:

$$L(p_{i,j}) = 4p_{i,j} - p_{i-1,j} - p_{i+1,j} - p_{i,j-1} - p_{i,j+1}.$$

The Laplacian is useful as a discontinuity-detection tool, *e.g.* for finding edges. The Laplacian of a point in an image with a locally uniform gradient is zero, even if the gradient magnitude at that point is high. We use the Laplacian transform to detect edges and the squared Laplacian values in an image region to measure the local smoothness of the image.

The Radon transform (Weisstein, 2003), which is related to the Hough transform, detects straight lines in an image. It requires its input image to be pre-transformed by, for example, an edge-detection algorithm, such that edges or discontinuities are highlighted. (We do so using a Laplacian-of-Gaussian convolution.) The Radon transform of an image $I(i, j)$ is an integral transform defined by

$$R(\theta, k) = \int_{-\infty}^{\infty} I(k\cos\theta - l\sin\theta, k\sin\theta + l\cos\theta) \, dl.$$

$R(\theta, k)$ is thus the integral of $I(i, j)$ along all points (i, j) on a line of orientation θ and of distance k from a parallel line passing through the origin $(0, 0)$. We use a discrete variant suitable for pixelated image data. We would interpret any large-magnitude $R(\theta, k)$ value as evidence of a straight edge in I oriented at an angle θ from the vertical with a distance k from a parallel line passing through the origin.

A notable characteristic of the Radon transform is that the straight lines it detects need not be composed of contiguous pixels: only the total number of pixels and their intensities matter. This characteristic is advantageous when analyzing an image with a sharp straight crystal edge periodically obscured along its length by dust, precipitate, skin or image noise. However, this same characteristic will confuse analysis when many unrelated discontinuities lie collinearly in the same image.

2.3.2. Detecting intra-droplet straight edges. The two straight-edge scores are based on computed evidence of the presence of straight lines inside the droplet. After eliminating the droplet border and the surrounding empty well from the image (see §2.2), the area inside the droplet is searched for straight edges by applying a Radon transform to the Laplacian of the Gaussian-blurred image. The Laplacian-of-Gaussian (LoG) operation suppresses high-frequency image noise and reveals edge pixels; the Radon transform, applied to edge data, detects straight lines.

From the result of the Radon transform, the first straight-edge score is obtained by summing the evidence weights of all significant straight lines detected, where a threshold for significance was empirically determined from an informal analysis of the Radon transforms of example images. The second straight-edge score is obtained from the maximum evidence weight of any straight line detected in the image.

2.3.3. Measuring local image smoothness. We use the Laplacian operator to measure image smoothness. First the droplet is subdivided into 16×16 pixel blocks and then the Laplacian values within each block are squared and summed. This yields a smoothness measure for each block. We select the highest score to use as our measure of image smoothness.

2.3.4. Detecting microcrystal features/microcrystal correlates. The microcrystal scores are based on comparisons of the target image with a library of ten microcrystal exemplars (selected by a crystallographer). Ten scores measure the presence of target-image features matching each exemplar and ten scores measure the presence of target-image features matching the inverse image of the exemplar.

The ten individual microcrystals were cropped from a set of example microcrystal-bearing images to produce a library of ten images between 12 and 18 pixels on a side (see Fig. 3).
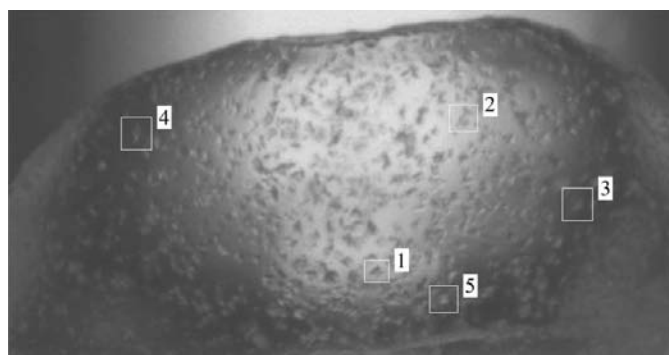

Figure 3
Microcrystal exemplars 1–5 and the surrounding regions of their source image.

Prior to cropping, the example images were transformed using the Laplacian operator.

20 microcrystal scores per target image were then generated by applying each microcrystal from the library as a linear filter against the Laplacian of the target image, effectively computing the correlation between every possible super-position of the Laplacians of the two images. Preliminary work on this feature-detection method revealed that strongly negative correlations are just as indicative of microcrystal presence as strongly positive correlations. An image feature correlating negatively with a microcrystal exemplar is equivalent to the same feature correlating positively with the inverse image of the same exemplar. A single microcrystal may appear as its inverse image under different lighting conditions. Thus, each exemplar is labeled with two correlation values, the maximum and the minimum.

Since the set of exemplars is limited, no individual microcrystal in a new image is likely to correlate strongly with any microcrystals in the library. This may not decrease the classification accuracy significantly, since microcrystals usually appear many at a time and thus at least one or two should correlate to an exemplar from the library.

2.3.5. Classification. Image classification is performed by linear discriminant analysis (LDA) using 23 numeric features that represent each image. Before any images may be classified, the linear discriminant must first be trained using a set of labeled images. Given two sets of 23-dimensional data points (X and N), LDA computes the vector $\mathbf{v}$ that maximally separates the points in X and N when projected onto $\mathbf{v}$. Further, LDA fits Gaussian distributions to X and N and calculates the distance along direction $\mathbf{v}$ that serves as the equiprobability boundary between the means of X and N. In essence, LDA produces a 23-element vector $\mathbf{v}$ and a numeric threshold t such that, given an image with feature vector $\mathbf{f}$,

$$\mathbf{f} \cdot \mathbf{v} > t \Leftrightarrow \text{the image is in class } X.$$

To classify a new image using LDA, we compute its feature vector $\mathbf{f}$ and the dot product $\mathbf{f} \cdot \mathbf{v}$ and then compare it with our threshold t.

3. Results

Our results are based on a set of 18 plates (27 648 images) hand-labeled by experts as either crystal-positive (X, 2224 images) or crystal-negative (N, 25 424 images). We regard the human expert labeling as ground truth.

3.1. Consistency

3.1.1. Machine consistency. We evaluated the consistency of our classification method by analyzing two plates containing independently run identical crystallization trials of glucose isomerase photographed after the same time lapse. A new linear discriminant was computed for each plate using LDA and training data drawn from the remaining hand-labeled plates. Shown below is a confusion matrix for the two plates, counting the number of images assigned to each class:

Machine classification (glucose isomerase)		Plate 2 N	X
Plate 1	N	923	123
	X	91	399

Thus, the machine classification for both plates agreed that 399 wells contained crystals and that 923 wells contained no crystals; they disagreed on the contents of the remaining 214 wells in each plate. The machine-assessed consistency rate was thus 86% for this experiment.

3.1.2. Human expert consistency and experimental repeatability. The same plates were also reviewed by a human expert:

Expert classification (glucose isomerase)		Plate 2 N	X
Plate 1	N	970	110
	X	84	372

Thus, the expert labeling of both plates agreed that 372 wells contained crystals and that 970 wells contained no crystals, while the remaining 194 wells showed conflicting results. These numbers yield a consistency rate and thus an experimental repeatability rate of 87% for this experiment.

3.1.3. Accuracy of the method on glucose isomerase trials. We then compared the human expert's assessment of each plate with the corresponding machine classification, summarized by these confusion matrices:

Glucose isomerase Plate 1		Machine classification N	X
Human expert	N	983	97
	X	139	317

Glucose isomerase Plate 2		Machine classification N	X
Human expert	N	981	73
	X	115	367

By assuming perfect accuracy of the human expert, we obtain accuracy scores[1] for our automated method: for plate 1, precision 0.77, recall 0.69 and accuracy 85%; for plate 2, precision 0.83, recall 0.76 and accuracy 88%.

3.2. Accuracy

To measure the overall accuracy of the system, we split our set of 18 expert-reviewed plates into training and test sets of equal size (each with 1112 X and 12 712 N images) and then

---

[1] Accuracy scores are derived from a confusion matrix which counts the number of true positives (TP, images correctly classified as bearing crystals), true negatives (TN, images correctly classified as non-crystal-bearing), false positives (FP, misclassified non-crystal-bearing images) and false negatives (FN, misclassified crystal-bearing images). Accuracy = (TP +TN)/(TP + TN + FP + FN); precision = TP/(TP + FP); recall = TP/(TP + FN).

Table 1
Effect on accuracy, precision, recall and ROC scores of removing features from analysis.

μX = microcrystal features (minimum and maximum correlations).

Excluded features	Δ accuracy (%)	Δ precision	Δ recall	Δ ROC
μX 1	−0.9	−0.015	−0.017	−0.005
μX 2	0	0	0.002	0
μX 3	−0.1	−0.001	0.003	−0.001
μX 4	−0.2	−0.004	−0.003	−0.002
μX 5	−0.4	−0.005	0.001	0
μX 6	−0.1	−0.001	−0.001	−0.001
μX 7	−3.5	−0.042	−0.006	−0.021
μX 8	0	0.001	0.005	−0.001
μX 9	−0.1	−0.002	0	0
μX 10	−0.5	−0.007	0.003	0
All μX	−8.9	−0.092	−0.053	−0.081
Radon sum	0	0	0	0
Radon max.	0	0	0.005	0.002
Both Radon scores	−0.3	−0.002	0.010	0.003
Image smoothness	−0.4	−0.004	0.007	0
All non-μX	−0.5	−0.005	0.008	0.001

Table 2
False positives.

Theme	No. of images	Comments
Speckled precipitate	92	>25% visually similar to microcrystals
Skin effects	20	
Contaminated wells	6	Hair or fibre
Mottled phase separation	1	
Genuine crystals	4	

performed feature extraction on each as described above. After computing the optimal linear discriminant using LDA on the training set, we evaluated the accuracy of the resulting classifier on the test set. The confusion matrix for our test set is shown below.
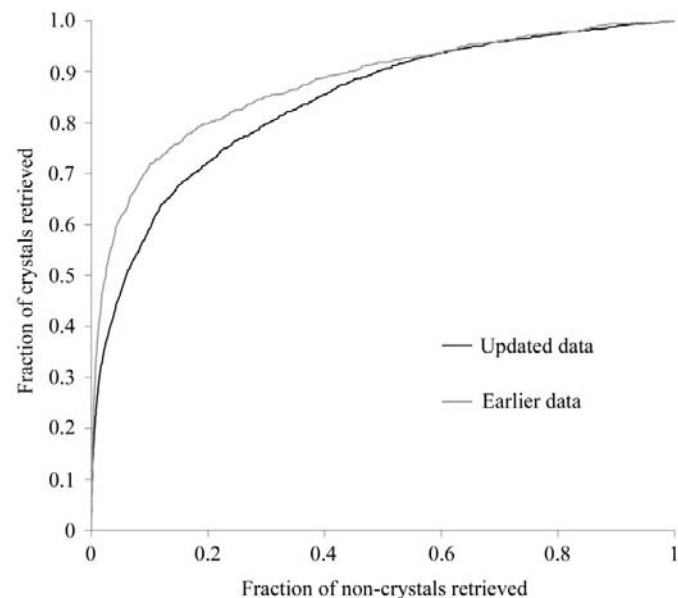
Test-set images		Machine classification	
		N	X
Human expert	N	11547	1165
	X	333	779

From this matrix, we obtain a test-set accuracy score of 89%. However, accuracy scores in this context may be deceiving. A 1536-well plate with 15 wells containing crystals may be classified with an accuracy exceeding 99% by a method claiming that no wells contain crystals. Thus, the crystal-detection task is better viewed as an information retrieval task. In retrieval terms, our method has a precision of 0.40 and a recall of 0.70.

3.2.1. Updated data. To test our method on a wider range of image data, we added an extra 14 plates (21 504 images) to our training and test sets, split equally between the two. A new linear discriminant was computed from the updated training set, and the performance of the classifier was evaluated on the updated test set:

Test-set images (updated)		Machine classification	
		N	X
Human expert	N	19842	3189
	X	525	1019

From this matrix, we obtain an updated test-set accuracy of 85%. In retrieval terms, our method has a precision of 0.24 and a recall of 0.66. These figures show a slight drop in classification accuracy, especially in the precision of the method. The higher scores of the original data set are explained in part by a skewed distribution of crystal forms. The population of crystal images in this set is dominated by crystals from the glucose
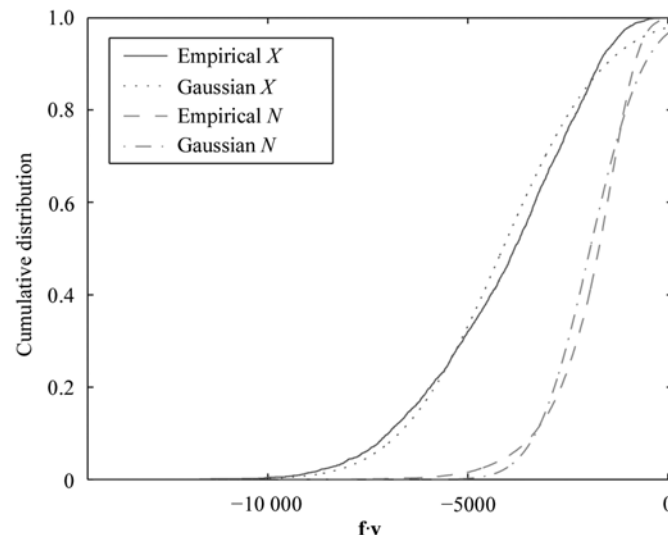

Figure 4
The ROC curves of image-retrieval exercises on two data sets: the original set of 18 plates (dominated by two very fruitful glucose isomerase plates) and an updated set with 14 extra plates added.

Figure 5
Empirical cumulative distributions of projected feature vectors of X and N images compared with their ideal Gaussian models.

isomerase plates. The updated data set contains a wider variety of crystal forms and a smaller proportion of crystals.
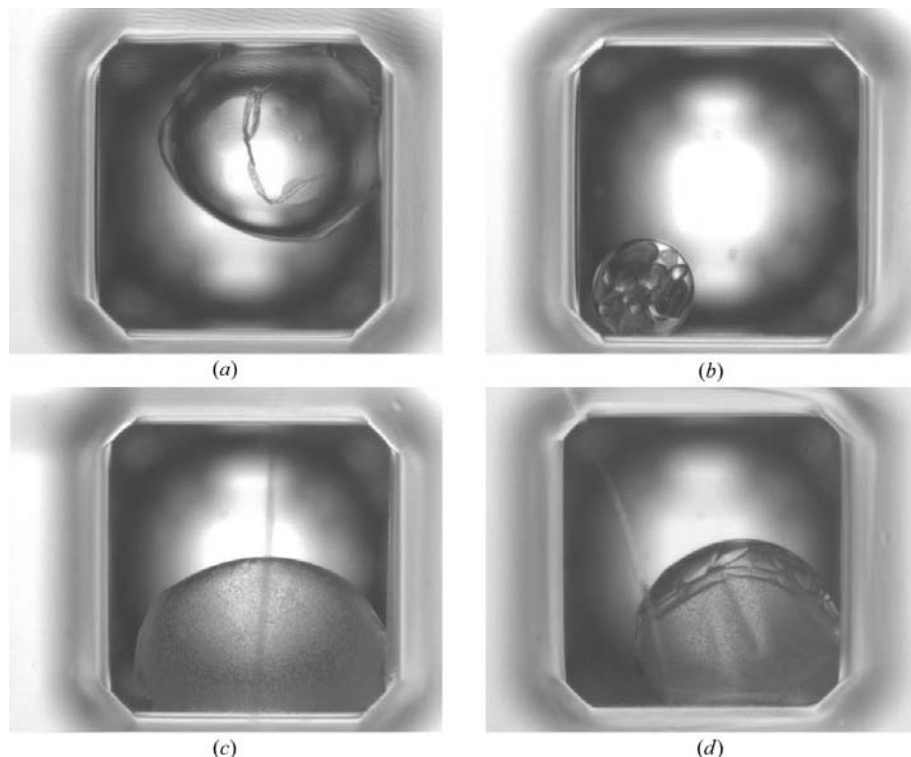

Figure 6
Misclassified non-crystals: (*a*) contamination, (*b*) salt crystals (dried-out drop), (*c*) speckled precipitate, (*d*) skin effects.
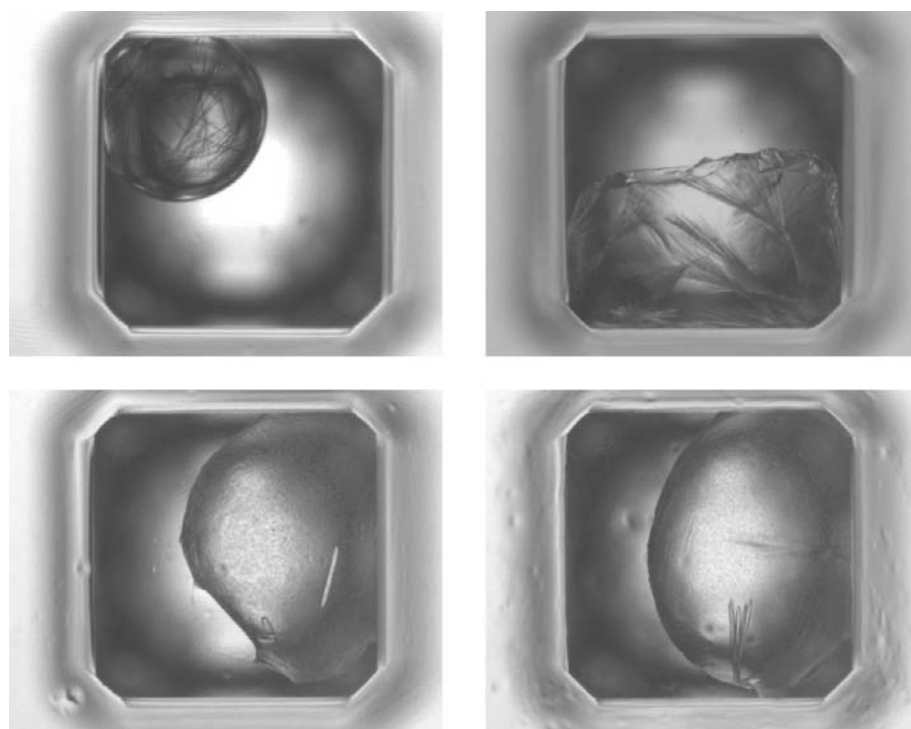

Figure 7
Genuine crystals detected by our method erroneously labeled as non-crystals.

3.2.2. Receiver operating characteristic. As described above, linear discriminant analysis of our training set established a linear combination of the 23 elements in the feature vector to produce a single numeric score that maximally (with respect to the training sets) discriminates between crystal-bearing and non-crystal-bearing images. LDA also produced the numeric threshold used to classify new images (*e.g.* the test sets). Choosing a threshold makes a trade-off between including as many crystal-bearing images while excluding as many non-crystal-bearing images as possible. Thus, the confusion matrices above and the precision and recall scores derived therefrom change as the threshold changes: greater precision at the expense of lower recall, or the reverse. The receiver operating characteristic (ROC) curves of these retrieval exercises illustrate this trade-off in Fig. 4. For the original retrieval exercise, the area under its ROC curve, called the ROC score, is 0.875. The ROC score of the retrieval on the updated data set is 0.84.

3.2.3. Impact of the various features. In order to measure the importance of the various features computed for each image, we re-analyzed the extended data set with various feature-dimensions removed. The changes in classification accuracy caused by these alterations are summarised in Table 1.

3.2.4. Suitability of Gaussian feature-modeling and linear discriminants. The linear discriminant analysis used here assumes Gaussian distributions of the feature vectors in the X and N sets. We chose a standard LDA because of its simplicity, small numbers of parameters and ultimate effectiveness. Formally, however, the Kolmogorov–Smirnov test rejects the hypothesis that either distribution is truly Gaussian. Fig. 5 plots the empirical cumulative distributions of the projected feature vectors ($\mathbf{f}\cdot\mathbf{v}$) for both X and N against the Gaussian distributions that model them, revealing asymmetrical tails. The same figure also reveals an overlap between X and N distributions: the root cause of misclassification.

3.3. Characterization of misclassified images

3.3.1. False positives. Study of the false positives generated by our method reveals distinct trends in the data. We manually reviewed 101 images falsely classified as bearing crystals. We noticed several themes present in these images, with one dominating (Table 2).

The classifier is apparently being fooled by grains of precipitate with a microcrystal-like appearance and by wrinkles in the skin that look like crystal edges (Fig. 6). The four genuine crystals (shown in Fig. 7) were confirmed as such, post-analysis, by HWI crystallographers.

3.3.2. False negatives. We reviewed 204 images falsely classified as non-crystal-bearing and again noticed several themes (Table 3).
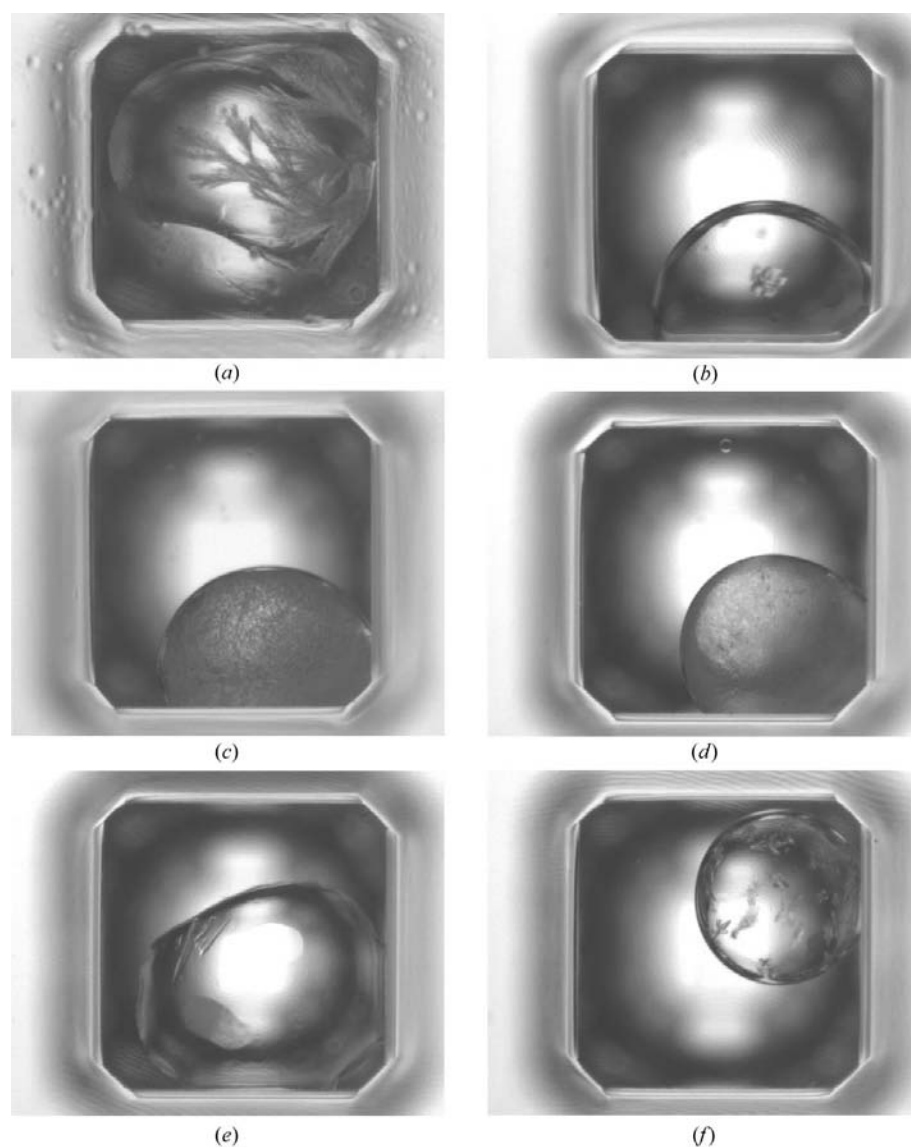
Here, the broader themes are (i) crystals too fine for detection and (ii) crystals without straight edges. Examples of these false negatives are shown in Fig. 8.

4. Discussion

The assessed accuracy of our method compares well with the results from an automatic classification of images from 96-well crystallization screens (Spraggon *et al.*, 2002; Wilson, 2002). In addition, our method achieves an accuracy comparable to the repeatability rate of the experimental setup. This means that a false result in a crystallization trial assessed by our method is as likely to arise from experimental noise as it is from classification error.

Despite the good results, we can learn from mistakes. Images misclassified by our system suggest several directions for improving classification accuracy.

(i) Our method misses the fine crystalline features of some images. It is likely that these features are being washed out by the noise-reducing Gaussian blur ($\sigma = 3$) applied to each image prior to the Radon transform step. It is also possible that the Radon intensity threshold is set too high. An improved method would apply multiscale analysis using a range of σ values and could calculate a histogram of Radon intensities rather than a count of threshold-exceeding Radon maxima.

(ii) The dendritic crystals described in §3.3.2 represent a significant fraction of misclassified images. We need to extract new features from our images that identify this feathery texture.

(iii) We must distinguish between microcrystals and speckled precipitates. Microcrystals typically have a sparkle (adjacent points of intensity significantly above and below the local mean intensity) that grains of precipitate do not have. An improved method would include a feature measuring this intensity difference.

We are currently extending the analysis method to classify the state of a single well at multiple time points. Knowledge of the state of a droplet at past or future time points can improve the accuracy of classifying the present time point (see Fig. 9).

In addition to improving accuracy, the method should be improved by broadening classification outcomes beyond the current $\{X, N\}$ dichotomy to include, for example, crystals, micro-
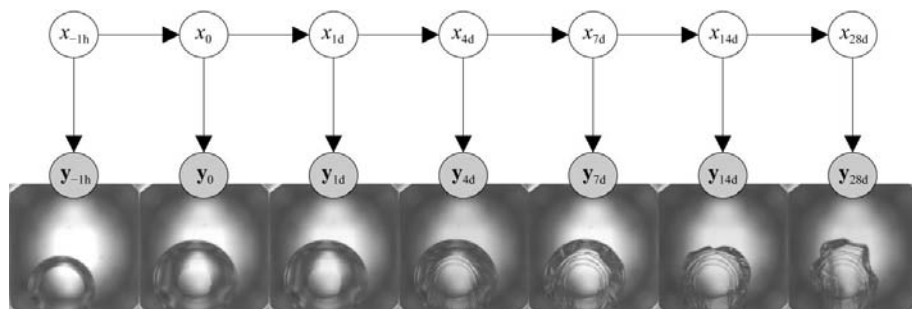

Figure 8
Misclassified crystals: (*a*) feathery dendrites, (*b*) out of focus, (*c*) multiple fine overlapping crystals, (*d*) precipitate-like microcrystals, (*e*) crystals in the poorly lit border region of the drop, (*f*) crystals in the poorly lit border region of the drop along with out of focus crystals in the illuminated portion of the drop.

Figure 9
Optimizing image classification by incorporating time-series data. A hidden Markov process models the changing state of the droplet contents in latent variables x. The observed variables y are the feature vectors computed from their respective images. The value of each x is inferred from the value of its corresponding y and its latent-variable neighbors.

Table 3
False negatives.

Theme	No. of images	Comments
Blurry crystals	17	
Feathery crystals	55	Crinkled, even sweeping, curvy feathery edges
Finely textured precipitate	22	
Piles of small crystals	28	Similar to those described by Wilson (2002)
Very fine small needle crystals	56	
No obvious crystal	3	Unusual drop textures
Crystals near droplet edge	9	Suspected segmentation error

(Jurisica, Rogers, Glasgow, Fortier *et al.*, 2001). The motivating hypothesis in each case is that proteins that respond similarly to a large set of crystallization trials will be similar in other characteristics. The method presented in this paper is crucial to establishing the database on which the CBR and data-mining systems depend, but has obvious benefits to crystallographers even without additional analysis as an automatic screen evaluation.

crystals, amorphous precipitate, skins, crystal aggregates (the 'crystal piles' of Wilson, 2002), phase separation and clear drops. Multiple objects within a single droplet could also be classified individually, in the manner of Wilson (2002), so that droplets containing different objects can be properly described.

One issue requires separate attention. Despite improved image-classification accuracy, we still do not have a reliable way of determining whether we are detecting salt or protein crystals. Our current and future work is focused on determining optimal dyes that could be added to the solution to enable differentiation of salt and protein crystals.

The image-classification system described here is one component of a larger intelligent system for reasoning on protein crystallization. There are two main goals of this system. The first is automated discovery, *via* data mining, of principles connecting protein-crystallization outcomes with data such as sample-preparation details and protein sequence, motif and family information. The second goal is the development of a case-based reasoning (CBR) system for protein crystallization. Its purpose is to assist in planning the search through the cocktail space for ideal crystal-growth conditions

References

Adams, J., Jewell, D., Jorgensen, K., Mickley, M. & Newman, J. (2002). *J. Assoc. Lab. Autom.* **7**, 36–40.

Bodenstaff, E., Hoedemaker, F., Kuil, M., de Vrind, H. & Abrahams, J. (2002). *Acta Cryst.* D**58**, 1901–1906.

Garyantes, T. K. (2002). *Drug Discov. Today*, **7**, 489–490.

Jurisica, I., Rogers, P., Glasgow, J., Collins, R., Wolfley, J., Luft, J. & DeTitta, G. (2001). *IEEE Intell. Syst. J.* **16**(6), 26–34.

Jurisica, I., Rogers, P., Glasgow, J., Fortier, S., Luft, J., Wolfley, J., Bianca, M., Weeks, D. & DeTitta, G. (2001). *IBM Syst. J.* **40**, 394–409.

Luft, J., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S. & DeTitta, G. (2001). *J. Cryst. Growth*, **232**, 591–595.

Murphy, K., Weiss, Y. & Jordan, M. (1999). *UAI'99: Proceedings of the 15th Conference on Uncertainty In Artificial Intelligence*, edited by K. B. Laskey & H. Prade, pp. 467–475. San Francisco: Morgan Kauffmann.

Rupp, B. (2003). *Acc. Chem. Res.* **36**, 173–181.

Spraggon, G., Lesley, S. A., Kreusch, A. & Priestle, J. P. (2002). *Acta Cryst.* D**58**, 1915–1923.

Weisstein, E. W. (2003). *Eric Weisstein's World Of Mathematics: Radon Transform*, http://mathworld.wolfram.com/RadonTransform.html.

Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.