Journal of
**Applied
Crystallography**

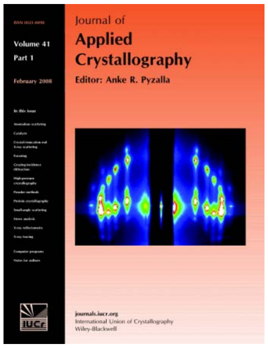
ISSN 0021-8898

Editor: **Anke R. Pyzalla**

Automated classification of crystallization experiments using wavelets and statistical texture characterization techniques

D. Watts, K. Cowtan and J. Wilson

J. Appl. Cryst. (2008). **41**, 8–17



Journal of
Applied
Crystallography

ISSN 0021-8898

Received 16 May 2007
Accepted 8 October 2007

Automated classification of crystallization experiments using wavelets and statistical texture characterization techniques

D. Watts,‡ K. Cowtan and J. Wilson*

Department of Chemistry, University of York, Heslington, York YO10 5YW, England. Correspondence e-mail: julie@ysbl.york.ac.uk

A method is presented for the classification of protein crystallization images based on image decomposition using the wavelet transform. The distribution of wavelet coefficient values in each sub-band image is modelled by a generalized Gaussian distribution to provide discriminatory variables. These statistical descriptors, together with second-order statistics obtained from joint probability distributions, are used with learning vector quantization to classify protein crystallization images.

1. Introduction

X-ray diffraction can provide the three-dimensional structure of biomolecules but relies on the growth of suitable crystals. Protein crystallization experiments are dependent on a number of parameters, including pH, temperature, precipitating agent and other additives. Identification of the optimal conditions for protein crystal growth is usually carried out by screening numerous combinations of these variables. High-throughput technology allows the automated screening of thousands of crystallization experiments a day in structural genomics centres, with imaging systems recording the results. Each experiment must be monitored over a period of time and the evaluation of large numbers of crystallization images by eye is tedious and time consuming. It is unlikely that software will ever replace the crystallographer's judgement completely, but automated categorization of the images according to experimental outcome would drastically reduce the number of images that require visual inspection.

Software to analyse crystallization images automatically, based on features that can be quantified and used for classification, is being developed by a number of research groups. Straight edges are an obvious characteristic of crystals that can be detected (Zuk & Ward, 1991). Spraggon *et al.* (2002) and Bern *et al.* (2004) use the Hough transform with Canny edge detection to calculate classification variables, while Cumbaa *et al.* (2003) use the Radon transform and compute an overall score from detected lines. Other features used for the classification of crystallization images include statistical descriptors, such as local variation (Bern *et al.*, 2004) and local smoothness (Cumbaa *et al.*, 2003), and global parameters obtained from local texture measures (Spraggon *et al.*, 2002; Saitoh *et al.*, 2005). The Fourier transform can readily identify periodic and directional structure and has been applied to extract textural information from crystallization images by Bern *et al.* (2004) and Walker *et al.* (2007). The approach of Pan *et al.* (2006) combines the use of texture features and intensity statistics with results from the Gabor wavelet decomposition.

Most methods for the analysis of crystallization images attempt to classify the crystallization drop as a whole, or as smaller blocks making up the drop. Currently, in the software *ALICE* (*Analysis of Images from Crystallization Experiments*), individual objects are identified within the crystallization drop and evaluated separately (Wilson, 2002). Objects are defined as connected sets of pixels above a threshold determined by the intensity statistics, so that faint, particularly out-of-focus, objects may not be identified and overlapping objects are not evaluated separately but as a conglomerate. Furthermore, the method does not take into account the relationships between objects or their spatial arrangement, and higher classification rates can be achieved by combining this approach with the use of textural information obtained from the crystallization drop as a whole.

Crystallographers can instantly assess the overall appearance of the crystallization drop, as well as considering individual objects within the drop. However, studies on the classification of crystallization images undertaken in both Oxford and York show surprisingly low agreement rates between crystallographers (unpublished work). In fact, images are often assigned to a different class when viewed by the same person on another occasion. The judgement can be subjective, so that accuracy and repeatability are not ensured, whereas automated image analysis, whilst lacking the power to comprehend the image scene entirely, can classify in a predictable and consistent manner. The variability in human classification must be taken into account when assessing the classification success of software to analyse images from crystallization experiments.

‡ Current address: EMBL Hamburg, Building 25A, Notkestrasse 85, 22603 Hamburg, Germany.

2. Imaging system

ALICE is being developed in collaboration with the Oxford Protein Production Facility (OPPF) at the University of Oxford where it is now used routinely to annotate images. The images shown were supplied by the OPPF, where crystallization experiments are performed in 96-well Greiner plates (micro-titre format) and the images are taken using an automated Oasis 1700 imaging system (Veeco, Cambridge, UK). Native images are 1024 × 1024 × 8-bit bitmap (BMP) images (∼1 Mbyte in size, corresponding to a pixel width of about 3 µm).

3. Wavelet decomposition

Various measures related to texture can be defined in terms of the correlation between intensities at different scales. As wavelet transforms decompose an image into different levels of detail, localized in both position and frequency, they are particularly suited to texture analysis. Pan *et al.* (2006) use the Gabor wavelet decomposition to provide eight sub-images from which they calculate a response for each pixel. Here, we use the multi-resolution analysis of Mallat (1989) with the simplest wavelet function due to Haar (1910). At each level of the wavelet transform, a combination of a low-pass filter, giving a smoothed approximation, and a high-pass filter, separating out the high frequency details, is applied. Applying the filters in both the x (horizontal) and y (vertical) directions gives four sub-images, each consisting of one-quarter of the pixels in the original image, for a single-level transform (Fig. 1). The vertical details in the image at this scale can be seen when the high-pass filter is applied along x and the low-pass filter along y, and *vice versa* for the horizontal details. Diagonal details are apparent in the sub-image obtained by applying the high-pass filter in both directions. Further levels of the transform are performed by applying the filters to the smooth approximation, *i.e.* the sub-image obtained by applying the low-pass filter in both directions. At level k of the transform, the number of pixels in each sub-image consists of $N/2^{2k}$ pixels, where N is the number of pixels in the original image.

4. Image pre-processing

In contrast with other authors, who have studied individual blocks within an image in order to classify texture (Cumbaa *et al.*, 2003; Bern *et al.*, 2004; Pan *et al.*, 2006), we have chosen to examine the texture of the whole crystallization drop for speed and simplicity. The image is first cropped to the size of the well and the drop identified on a coarse grid. This is achieved using a four-level wavelet transform to provide sub-images of approximately 2000 pixels, compared with the 700 × 750 pixels in the image of the entire well. The horizontal and vertical sub-images are then combined and thresholding applied to the resulting image to define a rough mask for the crystallization drop, as shown in Fig. 2. This mask is sufficient for the method involving classification of individual objects, as the drop edge and objects due to splatter outside the drop can be identified as such and ignored. However, a better mask is required for wavelet analysis, as the edge of the drop in particular would give rise to large wavelet coefficients. The mask is therefore refined by dilation until pixels identified as objects are reached. As objects outside the drop would prevent the mask reaching the drop edge (see Fig. 2*d*), this is performed after small objects beyond a certain distance from the centre of gravity of all object pixels have been deleted. Objects are defined as connected sets of pixels with gradient magnitude above a pre-defined threshold (Wilson, 2002). The gradient in the shadows around the drop satisfies this criterion, resulting in the halo around the drop seen in Fig. 2(*e*), which is easily removed by a further dilation of a few pixels (Fig. 2*f*).

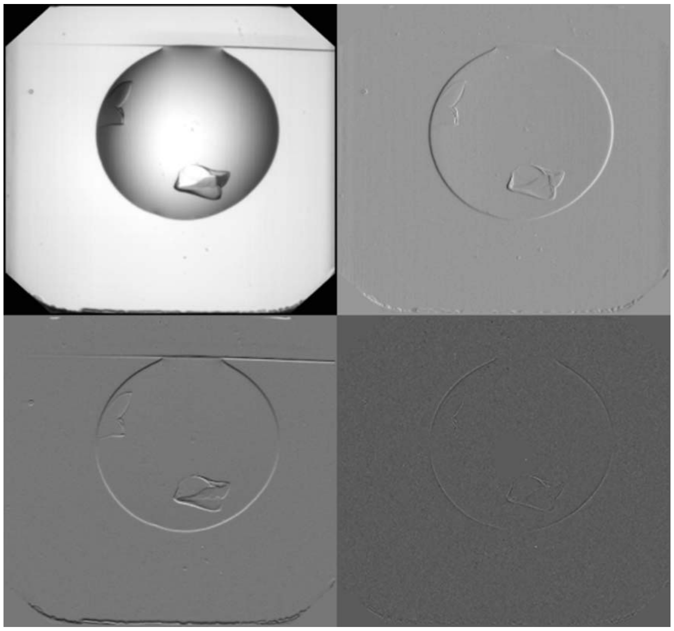


Figure 1
Single-level wavelet decomposition of a crystallization image. The four sub-images show smooth approximation, SXSY (top left), obtained by applying the low-pass filter in both x and y directions; the vertical details, DXSY (top right), obtained by applying the high-pass filter along x and the low-pass filter along y; the horizontal details, SXDY (bottom left), obtained by applying the low-pass filter along x and the high-pass filter along y; and the diagonal details, DXDY (bottom right), obtained by applying the high-pass filter along both x and y. Each sub-image has been rescaled independently to use the full range of display intensities.

5. Wavelet detail modelling

The wavelet coefficients (details) indicate change and are therefore zero for areas with no change and very small where the change in the image is small. The histogram of a detail image has a well defined shape with a single peak centred at zero (see Fig. 3). The many very small coefficients can be ignored, making wavelets ideal for image compression. Mallat (1987) considered the distribution of the wavelet coefficients in relation to the error on the reconstructed image and provided a mathematical model. It was shown that the histograms could be modelled by the functions

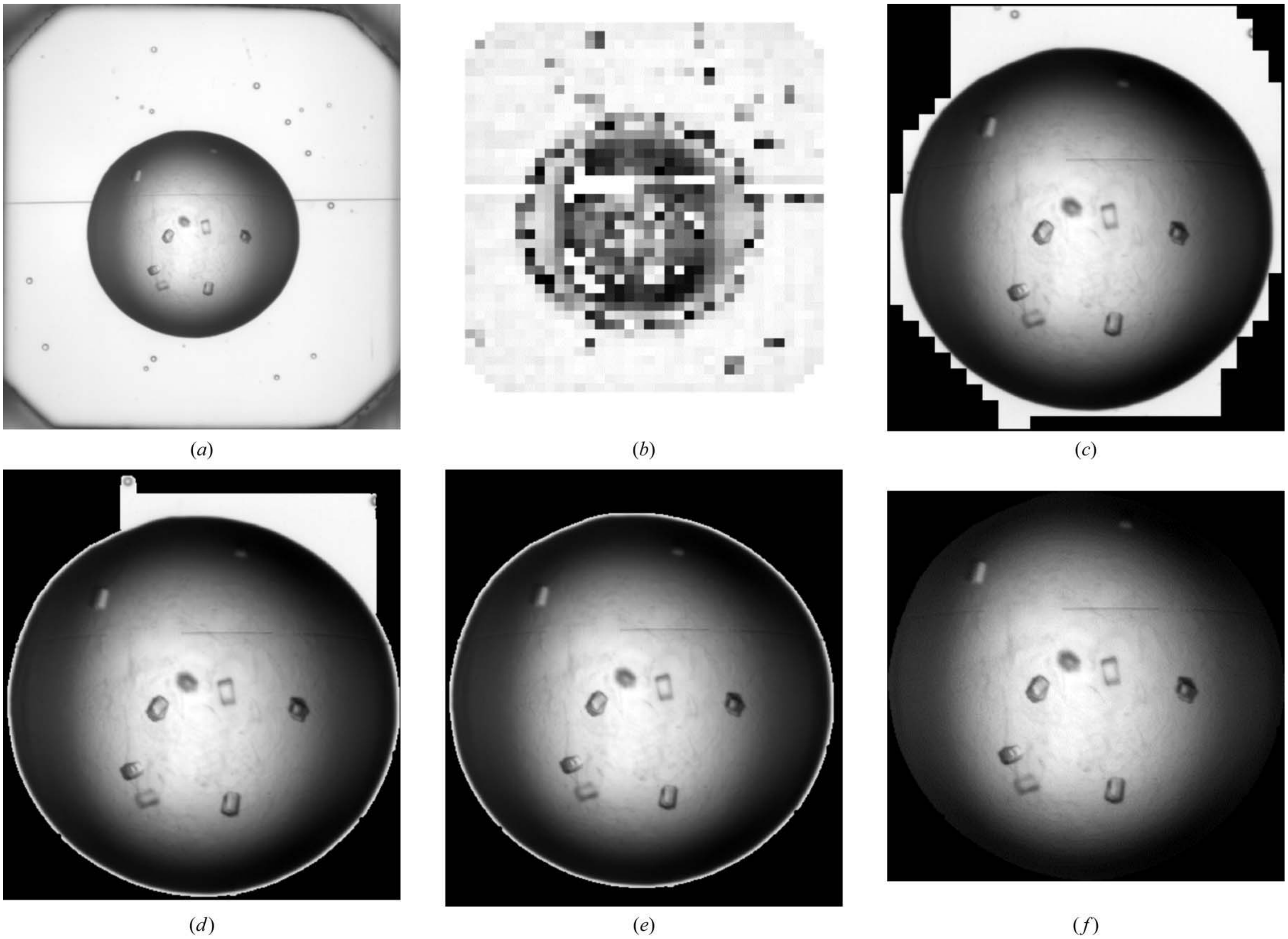


Figure 2
(*a*) The original image cropped to the size of the well. (*b*) The low-resolution image obtained from the combined fourth-level detail sub-images and (*c*) the rough mask obtained from this. (*d*) The problem of refining the mask in the presence of splatter is demonstrated and (*e*) resolved. (*f*) The final mask, after further dilation.

$$f(x) = \frac{N\beta}{2\alpha\Gamma(1/\beta)} \exp[-(|x|/\alpha)^{\beta}], \qquad (1)$$

where $\Gamma(x)$ is the gamma function defined by

$$\Gamma(x) = \int\limits_{0}^{\infty} t^{x-1} \exp(-t)\, \mathrm{d}t$$

and N is the number of pixels. Here, α models the variance and β the decay rate of the peak, with small values of β corresponding to a more pronounced peak shape. The family of functions includes the Gaussian distribution with $\beta = 2$ and the Laplacian distribution $\beta = 1$. The parameters α and β can be determined from the first and second moments of the histogram,

$$m_1 = \int\limits_{-\infty}^{\infty} |x| f(x)\, \mathrm{d}x, \qquad (2)$$

and

$$m_2 = \int\limits_{-\infty}^{\infty} x^2 f(x)\, \mathrm{d}x, \qquad (3)$$

as can be seen by substituting equation (1) in equations (2) and (3) and integrating. This gives

$$\alpha = \frac{m_1 \Gamma(1/\beta)}{\Gamma(2/\beta)} \qquad (4)$$

and

$$\beta = F^{-1}\left(\frac{m_1^2}{N m_2}\right), \qquad (5)$$

where

$$F(x) = \frac{\Gamma^2(2/x)}{\Gamma(3/x)\Gamma(1/x)}. \qquad (6)$$

In practice, we use

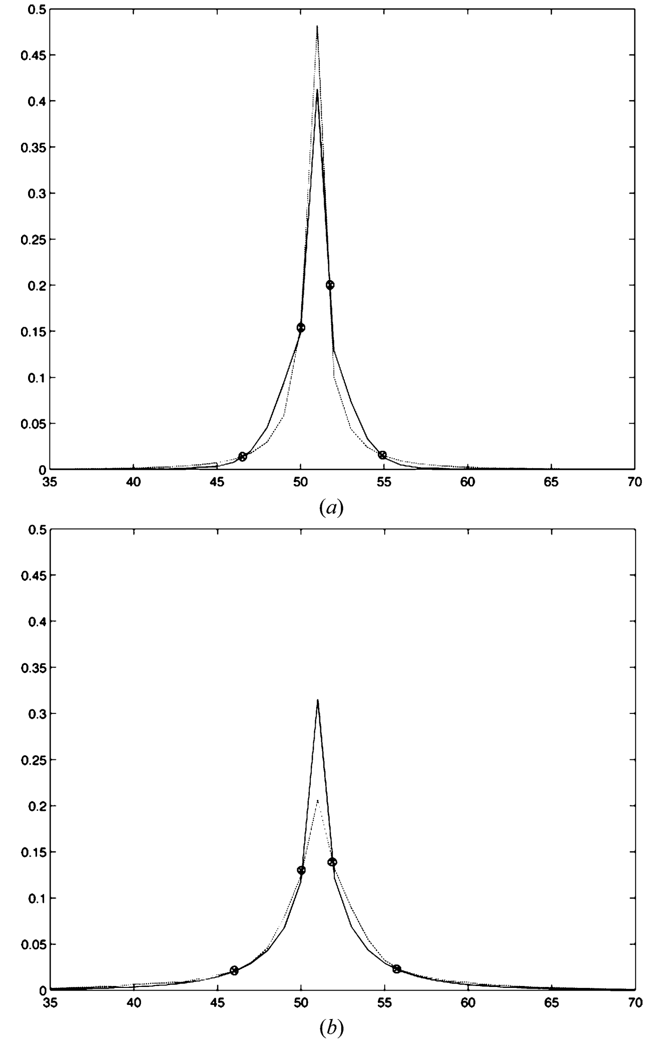


Figure 3
Histograms of detail images modelled by a generalized Gaussian. (*a*) The histogram of first-level detail coefficients (dotted line) and model (solid line) for an empty drop with α and β parameters of 1.24 and 1.14. (*b*) The histogram of first-level detail coefficients (dotted line) and model (solid line) for an image containing precipitate with α and β parameters of 1.80 and 1.47. The points defining the five different regions used to calculate 'lack-of-fit' parameters are marked on each histogram.

$$m_1 = \frac{1}{N}\sum_{i=1}^{N/2^j} \left|\omega_{j,i}\right| \tag{7}$$

and

$$m_2 = \frac{1}{N}\sum_{i=1}^{N/2^j} \omega_{j,i}^2, \tag{8}$$

where j denotes the level of the transform and $\omega_{j,i}$ is the ith wavelet sub-band coefficient at level j. The value of β was found using the technique described by Sharifi & Leon-Garcia (1995) with a look-up table of pre-calculated values for β between 0.01 and 7.0, in steps of 0.01. The inverse of the function in equation (5) is a monotonic increasing function so that, when the approximate value of β has been found from the look-up table, the value can be refined using the bisection method.

Van de Wouwer *et al.* (1999) (amongst others) applied this modelling technique for texture characterization and found that the statistics of the wavelet detail histograms could be used to classify natural textured images (bark, brick, fabric *etc.*). We have found that images from crystallization experiments can also be modelled by these generalized Gaussian functions and the parameters, α and β, used for classification. Fig. 3 shows the histograms and models obtained for an empty drop and for a drop containing precipitate. Empty drops give rise to many very small wavelet coefficients, leading to a narrow histogram with a very pronounced central peak, whereas drops containing precipitate give a wider spread of values and histograms with less sharp peaks. Such differences are reflected in the values of α and β.

We use the lack of fit to provide further discriminatory variables. In Fig. 3, the model for the empty drop is not quite as tall and narrow as the actual histogram, whereas the situation is reversed for the precipitate image. This situation is typical for images from these classes and can therefore be exploited for classification, along with other differences, such as a lack of symmetry in the histograms, which are often seen when crystals are present. In order to calculate further classification variables, the histograms were divided into regions by identifying four points (see Fig. 3). The first is the point where the model starts to increase more rapidly from zero, *i.e.* when $f(x) > 2/M$ for the first time, where M is the number of pixels in the wavelet sub-image. The symmetric equivalent gives the fourth point. The second and third points are either side of the central peak where the histogram and model first cross. Thus, five regions are obtained over which the difference between the histogram and the model can be calculated. In fact, it was found that the differences over the tails of the histograms could not be used for discrimination and so only three variables are calculated, these being the sum of the differences between the histogram and the model over the bins in each of the three regions of interest.

One further variable is calculated for each wavelet sub-image. This is the information entropy given by

$$H(x) = -\sum_{i=1}^{N} p(x_i)\log\left[p(x_i)\right], \tag{9}$$

where the probability, $p(x_i)$, is determined from the normalized histogram of wavelet coefficients for that sub-image. We apply a three-level wavelet transform giving three sub-images, corresponding to the horizontal, vertical and diagonal details, for each level, *i.e.* nine sub-images in total.

The algorithm to calculate classification variables from the wavelet detail coefficients can be summarized by the following steps:

(*a*) Perform a three-level two-dimensional wavelet transform to obtain nine detail sub-images.

Then, for each sub-image:

(*b*) Calculate the first and second moments using equations (7) and (8).

(*c*) Calculate α and β from equations (4) and (5).

(*d*) Obtain the histogram of wavelet coefficients and calculate the entropy using equation (9).

(*e*) Create the model using the parameters from step (*c*) and equation (1).

(*f*) Identify the four points that divide the histogram into different regions.

(*g*) For each of the three regions of interest, sum the differences between the histogram and the fitted model.

We therefore have six variables, α, β, entropy and the difference measures for three regions of the histogram, for each sub-image, giving a total of 54 potential classification variables. Univariate tests together with visual inspection were used in exploratory data analysis to assess the discriminatory power of the variables. All variables obtained from the first two levels of the wavelet transform appeared to provide some class separation. However, on level three, none of the variables for the horizontal or vertical sub-images appeared significant in *t*-tests, whereas, surprisingly, those for the diagonal details on this level did show significant discriminatory power. This was confirmed in classification, where better results were obtained using only 42 variables (six from each sub-image on each level except the third, where only the six variables from the diagonal sub-image were included) compared with those obtained when all 54 variables were used.

6. Joint wavelet statistics

As well as the first-order statistics determined from each detail sub-image, we also calculate statistics from joint conditional histograms. The decay of the wavelet coefficients across the levels of the transform can be used to characterize different types of edges (Mallet & Zhong, 1992). Sharp changes such as crystal edges give rise to large wavelet coefficients across the scales, whereas smoother changes in grey scale (due to shadows, for example) will produce wavelet coefficients that change gradually with subsequent levels of the transform. This information can be characterized by considering the correlation between corresponding wavelet coefficients on different levels of the transform.

The spatial and scale-to-scale dependency of wavelet coefficients has been utilized in image compression (Buccigrossi & Simoncelli, 1999) by modelling joint conditional histograms constructed between wavelet coefficient values at different levels. Pairs of wavelet coefficients at the same location in the wavelet sub-image are compared with a corresponding wavelet value in the same relative location at a neighbouring frequency scale. As we are interested in sharp features that may disappear at lower resolution, we condition the coarse-scale wavelet coefficients on those at a finer scale. It should be noted that this is the reverse of the approach of Buccigrossi & Simoncelli (1999), where the interest is image compression.

The second-order histogram is an estimate of the joint density function of pairs of wavelet coefficients. We use the absolute values of the wavelet coefficients, which are binned before the conditional histogram is created. In the joint histogram, the value at position (*i,j*) is the number of pixels with absolute wavelet coefficient value in bin(*i*) on the coarse scale and in bin(*j*) at the same location on the finer scale, *i.e.*

$$\mathrm{H}_n(i,j) = \sum_x \sum_y \begin{cases} 1, & \begin{array}{l} |\omega_k(x,y)| \in \mathrm{bin}(i) \ \text{ and} \\ |\omega_{k+1}(x,y)| \in \mathrm{bin}(j) \end{array} \\ 0, & \quad \text{otherwise}, \end{cases} \tag{10}$$

where ω_k are the wavelet coefficient values at level k and n indicates the sub-image (vertical, horizontal or diagonal details). Further notes on joint histograms are given in Appendix *A*. To increase the information content and provide some level of invariance to orientation, we combine the three detail sub-images to obtain

$$H_{i,j} = \sum_{n=1}^{3} H_n(i,j). \tag{11}$$

By normalizing the joint histogram $H_{i,j}$ so that the coarse-level (level $k + 1$) values sum to 1 for each fine-scale (level k) value, we obtain the joint conditional histogram $h_{i,j}$. This gives the probability that the coefficient value on level k lies in bin(*j*) given that the coefficient value at the corresponding position on level $k + 1$ lies in bin(*i*), *i.e.*

$$h_{i,j} = P\big[|\omega_{k+1}| \in \mathrm{bin}(j) \backslash |\omega_k| \in \mathrm{bin}(i)\big]. \tag{12}$$

Fig. 4(*b*) shows conditional histograms for second-level coefficients conditioned on the first-level coefficients for the three images in Fig. 4(*a*), where the conditioned wavelet values have been normalized so that brightness corresponds to probability. These plots were obtained from the original coefficient values rather than absolute values and show the characteristic bow-tie pattern first noted by Buccigrossi & Simoncelli (1999). In the central linear section of the pattern, the conditioned wavelet coefficients are close to zero, irrespective of the conditioning wavelet value. Outside this range, the distribution of the conditioned wavelet coefficients is dependent on the magnitude of the finer-scale coefficients. As the joint histograms are roughly symmetric about the horizontal and vertical axes, we use the joint conditional histogram of the absolute magnitudes in the statistical analysis, as shown in Fig. 4(*c*). It can be seen that the joint distributions differ between classes. For example, the variance of the conditioned coefficients for large values of the conditioning coefficients is greater for images containing many crystals and less for empty drops (see Fig. 4). Suitable statistical measures that describe the characteristics of the distributions can be exploited for classification.

For each distribution of the conditioning wavelet coefficients, *i*, we calculate the entropy using equation (9) and the expected value, variance, skewness and kurtosis according to the following equations:

$$E(i) = \sum_j h_{i,j} m_i, \tag{13}$$

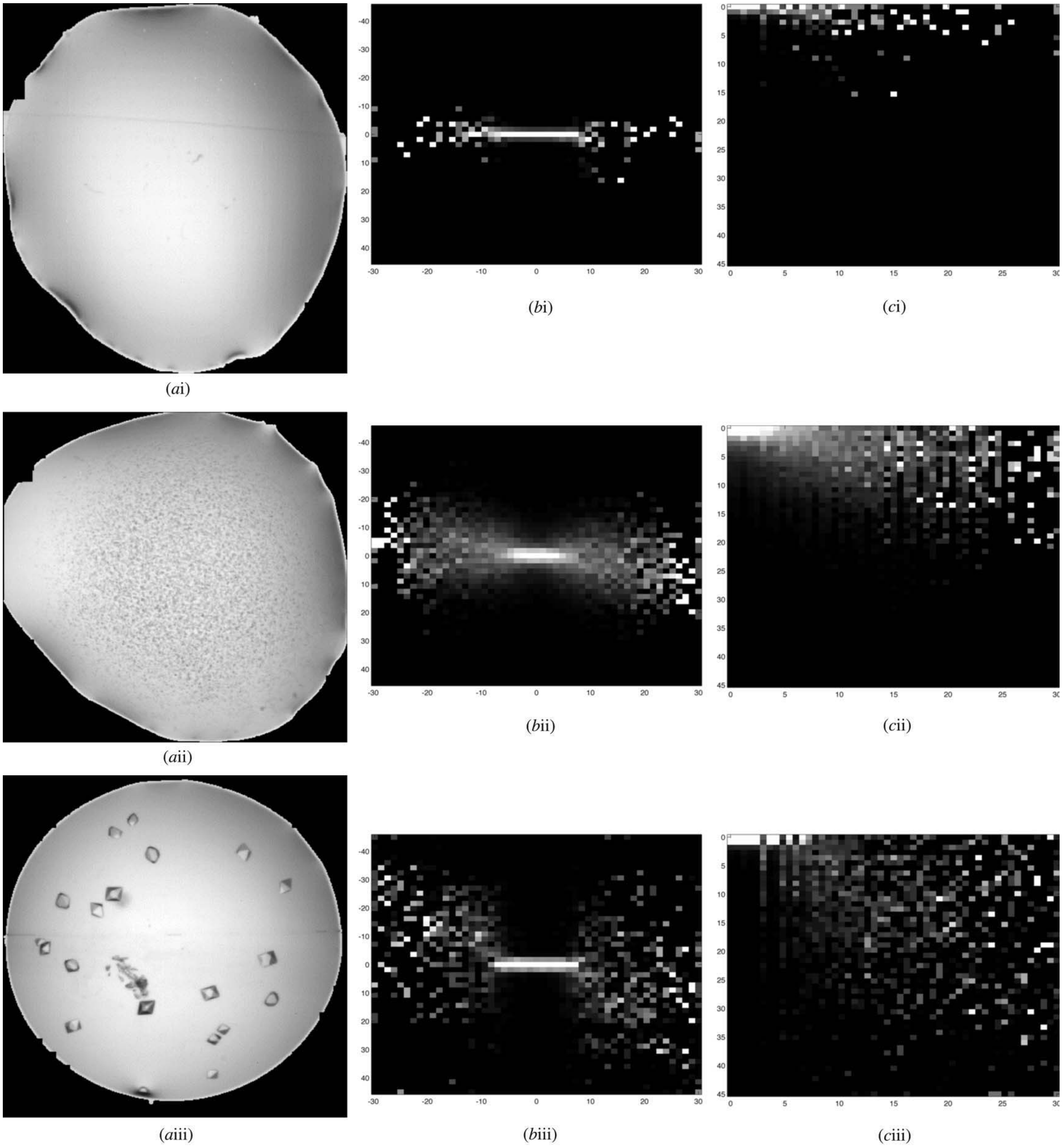
$$\mathrm{Var}(i) = E\Big\{\big[m_i - E(i)\big]^2\Big\}, \tag{14}$$

$$\mathrm{Skew}(i) = \frac{E\Big\{\big[m_i - E(i)\big]^3\Big\}}{\sigma^3}, \qquad (15)$$

$$\mathrm{Kurt}(i) = \frac{E\Big\{\big[m_i - E(i)\big]^4\Big\}}{\sigma^4}, \qquad (16)$$

where m_i is the midpoint of bin(i) and $\sigma = [\mathrm{Var}(i)]^{1/2}$.

Fig. 5(*a*) shows the expected value distribution for three typical images from each of the seven classes. For each image, the expected value of the absolute magnitudes of the conditioned coefficients is shown for each of the 31 bins of the conditioning wavelet coefficient histogram. Brighter pixels indicate higher values and it can be seen that different types of



(*a*i)

(*b*i)

(*c*i)

(*a*ii)

(*b*ii)

(*c*ii)

(*a*iii)

(*b*iii)

(*c*iii)

Figure 4

Conditional histograms from three images of different class. (*a*) Images of an empty drop, a drop containing precipitate and a drop containing crystals. (*b*) The conditional histograms obtained from these images for first-level wavelet coefficients conditioned on the second-level coefficients and (*c*) those using absolute values. Here, brightness corresponds to probability but has been rescaled to use the full range of display intensities. The *x* axis corresponds to the conditioning wavelet coefficient magnitude and the *y* axis to the conditioned wavelet coefficient magnitude.

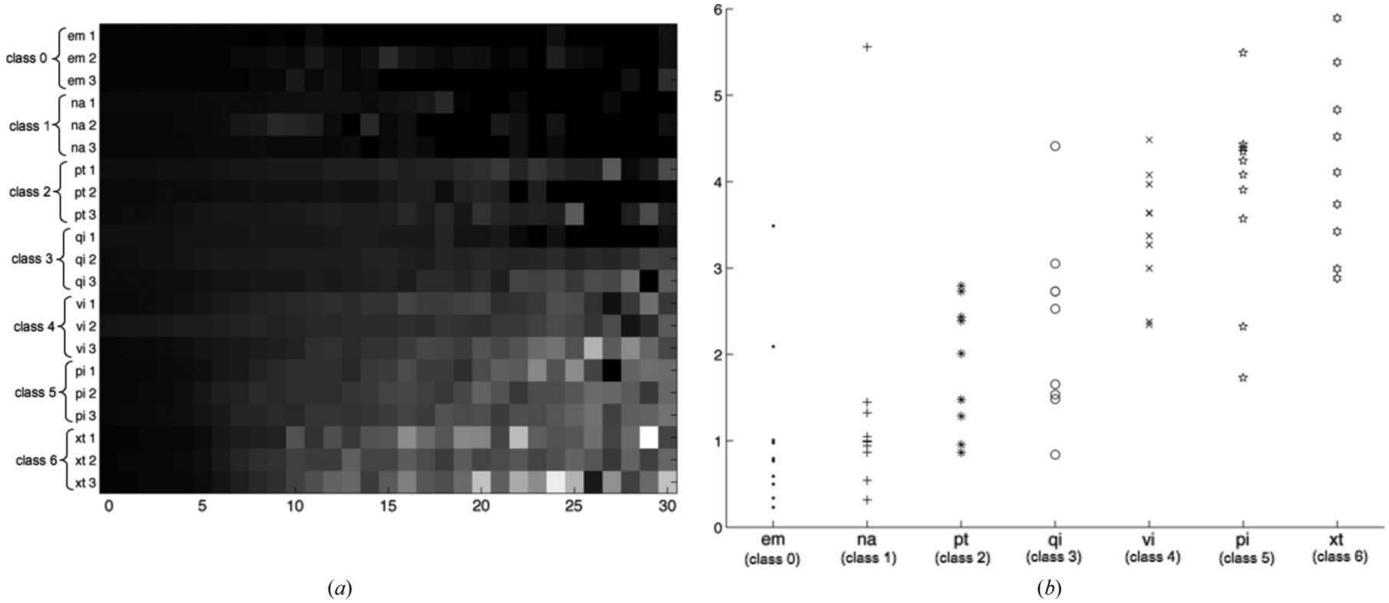


Figure 5
(*a*) The expected values calculated for each conditioning wavelet coefficient distribution, displayed as image intensities for three typical images from each of the seven classes. (*b*) The mean expected value for ten images from each class. Here, em indicates class 0, na class 1, pt class 2, qi class 3, vi class 4, pi class 5 and xt class 6.

image have different distributions of expected values. This is also true for the variance, skewness, kurtosis and entropy of the conditioned wavelet coefficient distributions.

The mean and the first to 99th percentile range of each of these statistical values over the histogram bins were found to summarize the distributions best and provide discriminatory variables. The average expected values for ten images from each class are plotted in Fig. 5(*b*), in which the correlation with class can be seen.

The algorithm to calculate classification variables from the joint conditional histograms can be summarized by the following steps:

(*a*) Obtain the absolute values of the wavelet coefficients for each of the nine detail sub-images from the three-level two-dimensional wavelet transform.

Then for $k = 1$ and $k = 2$:

(*b*) For each pair of sub-images (one from level k and one from level $k + 1$), calculate the joint histogram, $H_n(i, j)$, using equation (10).

(*c*) Combine the histograms from the three types of sub-image to obtain $H_{i,j}$ using equation (11).

(*d*) Normalize $H_{i,j}$ to obtain the conditional joint probability distributions, $h_{i,j}$.

(*e*) Calculate statistical measures for each conditioning distribution, i, using equations (9) and (13)–(16).

(*f*) Calculate the mean and percentile range of each statistical measure.

This gives a further 20 variables, *i.e.* both the mean and percentile range of each of the five statistical measures from the conditional joint probability distributions obtained from the first and second levels of the transform and from the second and third. Further pairwise combinations were considered but did not improve the classification results.

7. Results and discussion

Supervised learning algorithms are trained to associate a certain output with particular inputs so that the vector of values obtained from the classification variables, or feature vector, can be used to assign an object to a particular class. Training requires a set of input vectors that have been pre-classified by eye and an independent test set for validation to avoid problems with over-fitting. It should be noted that the number of classes is somewhat arbitrary, as crystallization experiments produce a continuum of results rather than discrete outcomes. Whatever the number, there will be some overlap between neighbouring classes, as demonstrated in studies to assess the reliability of human classification (unpublished work). Here, we use the following seven classes for crystallization outcomes:

0, empty drop;
1, denatured protein;
2, amorphous precipitate;
3, oil drops; phase separation; crystalline precipitate;
4, micro-crystals;
5, crystal clusters;
6, single crystals.

Fig. 6 shows an example image from each class. We found that this seven-class system gave an average agreement rate of ~70% for 16 crystallographers but that, in most cases where the scores do not agree, they differ by only one class. Allowing 'agreement' to tolerate a one-class difference gave an average agreement rate of ~94% and allowing a two-class difference gave close to total agreement, at ~99%.

For the wavelet method, descriptors were calculated for a training set of images classified by eye. This training set, consisting of 250 images from each of the seven categories, was used to train a learning vector quantization (LVQ) neural network (Kohonen *et al.*, 1992; the *LVQ_PAK* program can be downloaded from http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml). The resulting LVQ code vectors were used to classify an independent test set consisting of 150 images from each class. The training data for the object-based approach comprised ~9000 individual objects, classified by eye into 13 different object classes. The LVQ code vectors produced were used to classify the objects identified in the test set images and image scores were derived from these object scores.

Table 1 shows the results from both methods in the form of a truth table. The diagonal entries give the percentage of exact matches with the scores given when the images were classified by eye. However, misclassifications between neighbouring classes need not necessarily be considered incorrect if the continuous nature of the experiment and the diversity in human classification are taken into account. More serious misclassifications appear further away from the diagonal, with images classified too low above the diagonal and images classified too high below the diagonal.

The use of parameters derived using wavelet methods leads to a higher proportion of false positives (*i.e.* empty drops classed as crystals) and false negatives (crystals classed as empty drops) in comparison with the object-based method. This is not unexpected, as the majority of the crystallization drop can be empty in the case of single crystals, particularly those growing at the edge of the drop. In such cases, classification based on individual objects is likely to perform better. However, many images have features that occur throughout the drop and are better suited to texture-based methods. As the methods exploit complementary features, their combina-

Table 1
Agreement rates between the scores given by *ALICE* and those given when classified by eye.

The columns correspond to the results of automated analyses for each class (according to the visual score) in the rows of the table, so that the diagonal entries show exact matches. The figures given are percentages of the 150 images in each class (so that the values in each row sum to 100%) and bold type indicates acceptable classifications.

	6	5	4	3	2	1	0
Results obtained using classification variables calculated for individual objects within the drop							
Class 6	**41.3**	**26.7**	12.7	6.0	8.0	2.0	3.3
Class 5	**11.3**	**24.7**	**26.7**	14.0	14.0	8.0	1.3
Class 4	5.3	**12.0**	**34.7**	**19.3**	26.0	1.3	1.3
Class 3	2.0	4.0	**10.7**	**35.3**	**43.3**	1.3	3.3
Class 2	0.7	0.0	0.0	**9.5**	**89.2**	**0.7**	0.0
Class 1	1.4	3.4	4.7	7.4	**25.7**	**56.1**	**1.4**
Class 0	3.3	0.7	3.3	0.7	2.0	**9.3**	**80.7**
Results obtained using statistical measures obtained from wavelet analysis							
Class 6	**50.7**	**24.0**	15.3	3.3	0.0	0.0	6.7
Class 5	**20.0**	**38.0**	**34.7**	3.3	2.7	0.0	1.3
Class 4	16.0	**24.0**	**20.7**	**14.7**	13.3	6.0	5.3
Class 3	5.3	6.0	**12.7**	**33.3**	**28.7**	10.0	4.0
Class 2	0.7	0.0	3.4	**4.7**	**83.8**	**7.4**	0.0
Class 1	2.0	6.8	14.2	7.4	**12.8**	**54.7**	**2.0**
Class 0	10.7	1.3	0.0	1.3	2.0	**0.7**	**84.0**

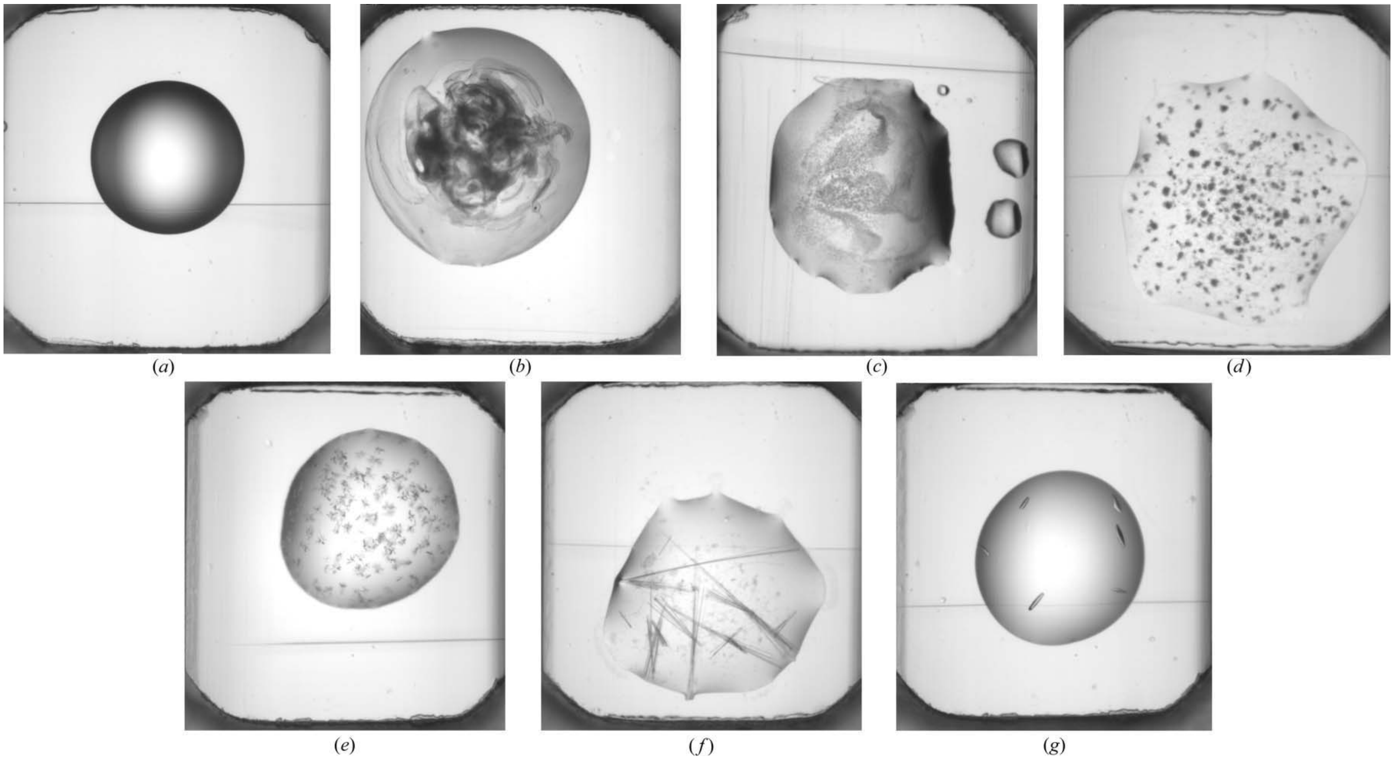


Figure 6
Example images from the seven categories used in classification. (*a*) Empty drops score 0. (*b*) Drops with denatured protein or other unwanted outcomes score 1. (*c*) Precipitate scores 2. (*d*) Outcomes better than precipitate but non-crystalline score 3. (*e*) Micro-crystals score 4. (*f*) Crystal clusters score 5. (*g*) Single crystals score 6.

Table 2
Classification rates for the scores given by *ALICE* when the object-based method and the texture-based approach are combined.

The columns correspond to the results of automated analyses for each class (according to the visual score) in the rows of the table and are given as percentages of the 150 images in that class. Bold values indicate acceptable classifications. The results were obtained by taking the maximum score from the two individual methods.

	6	5	4	3	2	1	0
Class 6	**61.3**	**23.3**	8.7	2.0	0.7	0.7	3.3
Class 5	**23.3**	**37.3**	**32.0**	4.7	1.3	0.0	1.3
Class 4	18.0	**24.7**	**32.7**	**16.0**	6.7	0.7	1.3
Class 3	5.3	7.3	**18.7**	**45.3**	**21.3**	0.0	2.0
Class 2	1.4	0.0	3.4	**13.5**	**81.1**	**0.7**	0.0
Class 1	2.7	8.8	15.5	12.2	**21.6**	**38.5**	**0.7**
Class 0	7.3	1.3	2.7	0.7	2.0	**8.7**	**77.3**

Table 3
Summary of classification results.

The three columns correspond to automated scores indicating images that should definitely be looked at by the crystallographer (yes), those that maybe should be looked at (maybe) and images that need not be looked at (no). Each row of the table shows the class according to the visual score, where 'crystals' includes classes 5 and 6, 'micro-crystals' corresponds to class 4, 'precipitate' includes classes 2 and 3, and 'rubbish/empty' covers classes 0 and 1.

	Yes	Maybe	No
Results obtained by the object-based method			
Crystals	71.7	21.0	7.3
Micro-crystals	52.0	45.3	2.6
Precipitate	8.7	88.7	3.6
Rubbish/empty	8.4	17.9	73.8
Results obtained by the wavelet-based method			
Crystals	91.4	4.7	4.0
Micro-crystals	60.7	28.0	11.3
Precipitate	14.1	75.3	10.7
Rubbish/empty	17.5	11.8	70.7
Results obtained when the two methods are combined			
Crystals	93.0	4.4	2.7
Micro-crystals	75.4	22.7	2.0
Precipitate	18.1	80.6	1.4
Rubbish/empty	19.2	18.3	62.6

tion leads to improved classification rates. In crystallization image categorization, false negatives mean missing crystals and are therefore much more serious than false positives. The final score for the image is taken as the maximum score obtained by either method individually and the results are shown in Table 2. However, in order to reduce the number of false positives, in cases where the wavelet method indicates crystals and the object-based method suggests an empty drop, the image score is taken to be zero.

As truth tables involving seven classes can be difficult to analyse, the results from Tables 1 and 2 are summarized in Table 3. Here, the visual scores are represented by just four classes and the automated scores by just three: images that should definitely be looked at by the crystallographer (yes), those that maybe should be looked at (maybe) and images that need not be looked at (no). The relatively large number of false positives occurring is due mainly to images from class 1, which contains examples with denatured protein, wrinkled skin across the drop and foreign bodies such as fibres. Such images will invariably look more interesting in terms of the classification variables than they really are and more importance is given to reducing the number of false negatives, *i.e.* missed crystals. It should be noted that our test set was not carefully chosen, but rather contains real examples with problems due to focus, light and shadows. Fig. 7 shows two of the images containing crystals that were classified incorrectly. In Fig. 7(*a*), the misclassification is due to the very low contrast and, in Fig. 7(*b*), it is due to the fact that the crystals lie in shadow at the edge of the drop. Work to reduce the number of false positives due to such problems will continue.

Rather than determining the class for each object or image from a single winning vector, the LVQ can be used to provide a probability for every class, calculated from the classes of the top ten vectors weighted according to the distance measure. A better way of combining the results using probabilities is being

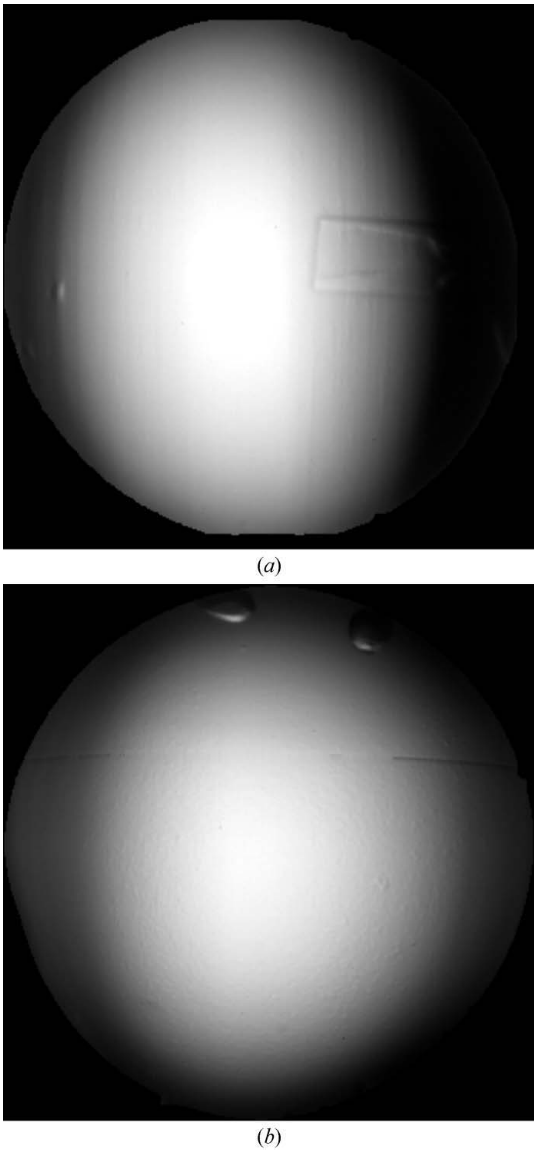


Figure 7
Examples of images that were misclassified. Although both images contain crystals, the poor contrast in (*a*) and the shadows at the edge of the drop in (*b*) caused these images to be classed incorrectly.

sought and will ideally also allow the class probabilities of one object to influence those of another.

The aim is not to replace human classification but to reduce the number of experiments that need to be examined by eye. Automated scoring allows the images to be examined in order of merit and, when high-scoring conditions are confirmed, no further images need be considered. The potential for the development of optimization procedures is also important; not only will many promising initial conditions be identified, but failed experiments will also be recorded, allowing the possibility for automated screening protocols to be developed *via* data-mining techniques.

APPENDIX *A*

Co-occurrence matrices provide a means of quantifying correlated features between different images. The joint histogram is a special case of a co-occurrence matrix where the values are binned into non-overlapping intervals. The co-occurrence matrix gives the frequency of occurrence of a pixel value in one image for any given value of the same pixel in the other image. For example, consider two simple 5×5 pixel images, *A* and *B*, whose pixels can take four values, 0–3.

A

0	1	2	3	0
1	2	3	0	1
2	3	0	1	2
3	0	1	2	3
0	1	2	3	0

B

2	2	3	1	3
0	0	1	2	2
3	1	2	0	0
1	0	3	3	1
3	2	0	1	0

The co-occurrence matrix of the *A* and *B* matrices is the 4×4 matrix shown below (since there are four possible pixel values). The elements in the matrix are the number of times a given value in *A* leads to a given value at the corresponding pixel in *B*.

***B*\A**	***0***	***1***	***2***	***3***
0	2	2	3	0
1	0	0	0	6
2	3	3	3	0
3	2	1	0	0

From this we can immediately see that, while the other values are fairly uncorrelated, a 3 in *A* always leads to a 1 in *B*, and a 1 in *B* always leads to a 3 in *A*.

In the current problem, wavelet coefficients at one level of detail may be correlated in different ways with coefficients at another level. Individual pixel noise and fine-grained textures are likely to be uncorrelated between detailed and coarse images, whereas strong edges will give rise to features across several levels of detail.

The authors would like to thank the BBSRC for funding this research. JW is a Royal Society University Research Fellow.

References

Bern, M., Goldberg, D., Stevens, R. C. & Kuhn, P. (2004). *J. Appl. Cryst.* **37**, 279–287.

Buccigrossi, R. W. & Simoncelli, E. P. (1999). *IEEE Trans. Image Proc.* **8**, 1688–1701.

Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J. R., DeTitta, G. & Jurisica, I. (2003). *Acta Cryst.* D**59**, 1619–1627.

Haar, A. (1910). *Math. Ann.* **69**, 331–371.

Kohonen, T., Kangas, J., Laaksonen, J. & Torkkola, J. (1992). *Neural Netw.* **1**, 725–730.

Mallat, S. (1987). *Proc. IEEE Workshop Comput. Vis.* Miami, Florida, USA, pp. 2–7.

Mallat, S. (1989). *IEEE Trans. Patt. Anal. Mach. Intell.* **11**, 674–693.

Mallet, S. & Zhong, S. (1992). *IEEE Trans. Patt. Anal. Mach. Intell.* **14**, 710–732.

Pan, S., Shavit, G., Penas-Centeno, M., Xu, D.-H., Shapiro, L., Ladner, R., Riskin, E., Hol, W. & Meldrum, D. (2006). *Acta Cryst.* D**62**, 271–279.

Saitoh, K., Kawabata, K., Asama, H., Mishima, T., Sugahara, M. & Miyano, M. (2005). *Acta Cryst.* D**61**, 873–880.

Sharifi, K. & Leon-Garcia, A. (1995). *Proc. IEEE Trans. Circuits Syst. Video Technol.* **5**, 52–56.

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

Van de Vouwer, G., Scheunders, P. & Van Dyck, D. (1999). *IEEE Trans. Image Process.* **8**, 592–598.

Walker, C. G., Foadi, J. & Wilson, J. (2007). *J. Appl. Cryst.* **40**, 418–426.

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

Zuk, W. & Ward, K. (1991). *J. Cryst. Growth*, **110**, 148–155.