

# Recognition of Micro-Array Protein Crystals Images using Multi-scale Representations

Ya Wang, David H. Kim, Elsa D. Angelini\*, Andrew F. Laine

Heffner Biomedical Imaging Laboratory,  
Department of Biomedical Engineering, Columbia University, NY  
\* Ecole Nationale Supérieure des Télécommunications  
Département Traitement du Signal et des Images (TSI), Paris, France

## Abstract:

Micro-array protein crystal images are now routinely acquired automatically by CCD cameras. High-throughput automatic classification of protein crystals requires to alleviation of the time-consuming task of manual visual inspection. We propose a classification framework combined with a multi-scale image processing method for recognizing protein crystals and precipitates versus clear drops. The main two points of the processing method are the multi-scale Laplacian pyramid filters and histogram analysis techniques to find an effective feature vector. The processing steps include: 1. Tray well cropping using Radon Transform; 2. Droplet cropping using an ellipsoid Hough Transform; 3. Multi-scale image separation with Laplacian pyramidal filters; 4. Feature vector extraction from the histogram of the multi-scale boundary images. The feature vector combines geometric and texture features of each image and provides input to a feed forward binomial neural network classifier. Using human (expert crystallographers) classified images as ground truth, the current experimental results gave 86% true positive and 94% true negative rates (average true percentage is 90%) using an image database which contained over 2,000 images.

To enable NESG collaborators to carry our crystal classification, a web-based Matlab© server was also developed. Users at other locations on the internet can input micro-array crystal image folders and parameters for training and testing processes through a friendly web interface. Recognition results are shown on the client side website and may be downloaded by a remote user as an Excel© spreadsheet file.

Keywords: Multi-scale, Laplacian pyramid, protein crystals, neural network, Radon Transform, Hough Transform, microscopic images

## 1. Introduction

Recent development and testing of a multi-scale image analysis technique for automatically classifying protein precipitates and crystals based on microscopic images is presented. High-throughput (HTP) protein images are acquired by a CCD camera under robotic control, which captures the contents within each micro-array tray well on a 1536-wells matrix plate. The aim is to find all wells that contain droplets with protein crystals or precipitates at an early stage. Binary classification will then designate well images with crystals or precipitates as "1" and clear drops as "0." The principle issues in this classification problem are the non-uniform lighting conditions, irregular droplet boundaries, and varying droplet size. Using Radon and Hough transforms [1], the nonessential fringe areas of the image are cropped out and the remaining image is decomposed into a multi-level Laplacian pyramid [2]. Each level of the Laplacian pyramid has a different resolution and contains distinct features; only higher levels which contain most of the boundary information are used to calculate the statistical parameters included in our feature vector representation for image classification. Fast algorithms are used in our coding to meet a real-time, web-based remote classification requirement. Feature vectors are composed of 8 parameters which describe the histogram characters [3] of each level of Laplacian pyramidal expansion. Classification is accomplished by feeding the feature vectors to a linear or nonlinear feed forward neural network [4]. Expert-labeled images are used as ground truth to compare with our classification results. Our recent tests have yielded a high classification accuracy of about 90% or higher with longer computational time. Further refinements are discussed at the end of this paper. Our previous work was introduced in [5].

## 2. Methods

### 2.1 Well cropping with Radon Transform

Unlike previous approaches that first attempt to delete the background using a reference border image that contains an empty “model” well, we first apply a Radon Transform for cutting out the well from the larger field of view. Figure 1 shows an original micro-array genomic image and the Radon Transform of this image with four peaks representing the four main well boundaries. The peak locations can be easily found and from which we can calculate the x, y location of the well boundaries. Thus the well can be cropped out accurately. Results using the background image directly are compared with Radon transform in Figure 2. Improvements obtained using the Radon Transform cropping method are obvious in Figure 2. Since each image may have a slightly different shifting, it is more accurate to use the Radon Transform for well cropping.

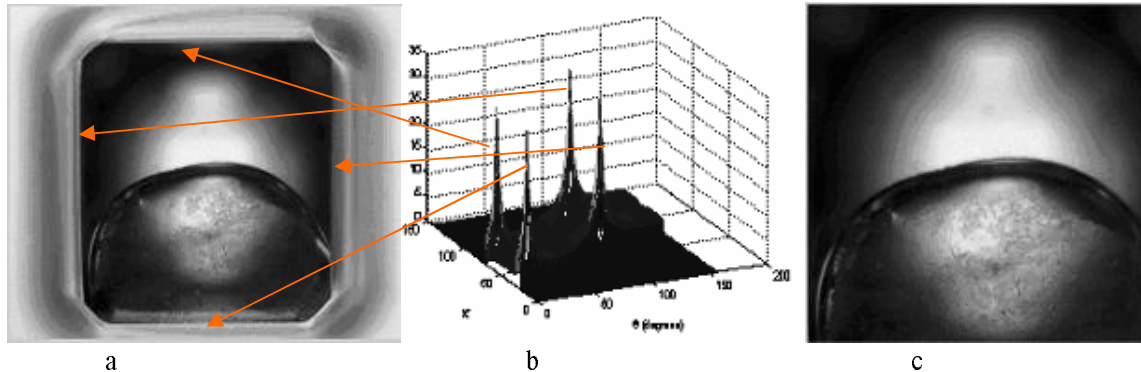


Figure 1: a. Micro-array genomic image. b. Radon Transform of the image in a. c. Cropped image with Radon Transform.

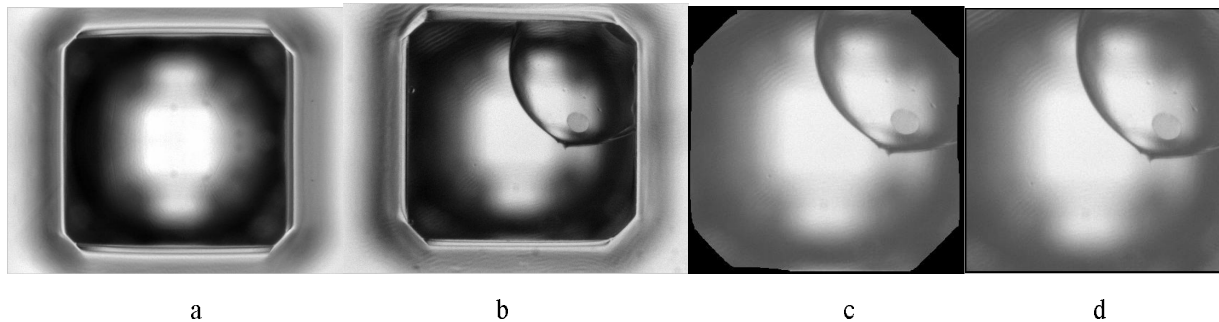


Figure 2: a, Background image. b, Original image. c, Image cropped according to well contour of the border image. d, Image cropped with Radon transform.

## 2.2 Droplet cropping with an Ellipsoidal Hough Transform

An ellipsoid Hough Transform algorithm was used to identify ellipse objects in an image, based on edge maps. Three most probable ellipses are selected to find the Region of Interest (the oil bubble that may contain protein crystals) inside the well region. To reduce the computational cost of this algorithm, we followed a modification of the algorithm as suggested in [6] which reduces the search to a three-dimensional space by exploiting the gradient information of the image. Figure 3 below shows the resulting well-cropped image, the three most probable ellipses found by the Hough Transform (shown as Red, Green and Blue lines), and the final cropped ROI, to be further processed.

## 2.3 Laplacian Pyramid Filter

Multi-scale Laplacian pyramid technique were originally used in image encoding [2]. Each level of the Laplacian pyramid is the prediction error  $L$  which is given by subtracting a low-pass filtered image from the original image. Then the low-pass filtered image may be represented at reduced sample density. Iteration of the two processes will generate a pyramid like image series as shown in Figure 4.

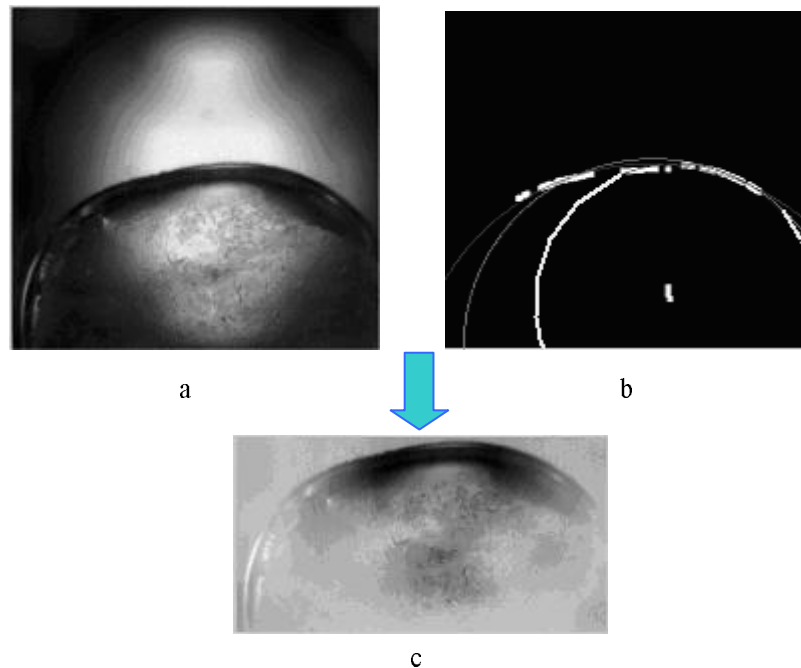


Figure 3: ROI cropping with Hough Transform. a, Image with well cropped using Radon Transform. b, Three ellipses detected by the Hough Transform. c, Rectangular area containing the ROI.

Here Laplacian pyramid filters are used to extract boundary (gradient) information from the cropped images. The code elements are localized in spatial frequency as well as in space [2]. Three-level decomposition is selected for feature extraction and representation according to experimental tests. The motivation for using multi-scale filters is to extract useful information contained in the original images to keep the size of the feature vector, input of the neural network, small.

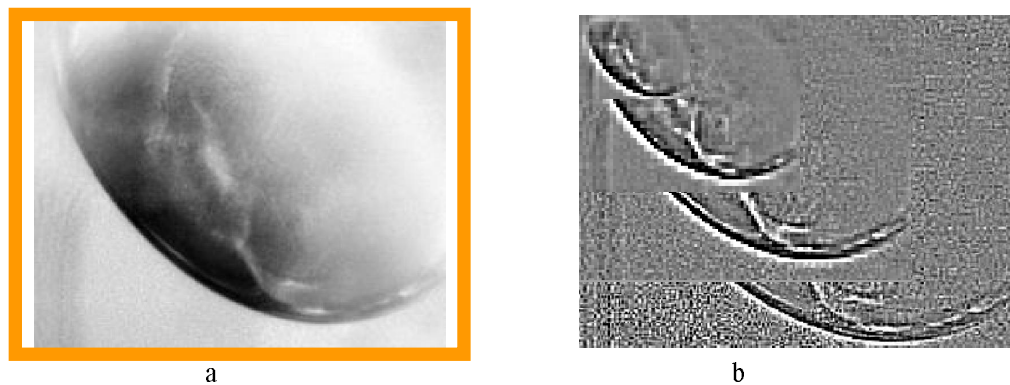


Figure 4: Boundary information extraction with Laplacian Pyramid. a, Original image. b, Laplacian pyramid series.

#### 2.4 Feature extraction and representation

The feature vector contains quantitative shape descriptions of the first and second-order histogram of Laplacian pyramid coefficients. This feature vector provides a more complete representation of the data driving the neural network as input. The shape of an image histogram provides many clues as to the character of the protein crystal image. The selected quantitative shape descriptors of a first-order histogram are:

Mean: 
$$S_M \equiv \bar{b} = \sum_{b=0}^{L-1} bP(b)$$

Standard Deviation: 
$$S_D = \left[ \sum_{b=0}^{L-1} (b - \bar{b})^2 \right]^{1/2}$$

Skewness: 
$$S_S = \frac{1}{\sigma_b^3} \sum_{b=0}^{L-1} (b - \bar{b})^3 P(b)$$

Kurtosis: 
$$S_K = \frac{1}{\sigma_b^4} \sum_{b=0}^{L-1} (b - \bar{b})^4 P(b) - 3$$

Energy: 
$$S_N = \sum_{b=0}^{L-1} [P(b)]^2$$

Entropy: 
$$S_E = - \sum_{b=0}^{L-1} P(b) \log_2 [P(b)]$$

One of the second-order histogram features is used:

Autocorrelation: 
$$S_A = \sum_{a=0}^{L-1} \sum_{b=0}^{L-1} abP(a, b)$$

and

Power: 
$$S_P = \sum_{Im}^{L-1} FFT(Im)^2$$

Where  $P(b)$  is the first-order histogram estimate.  $P(a, b)$  represents the histogram estimate of the second order distribution. Parameter  $b$  is the pixel amplitude value.  $L$  is the upper limit of the quantized amplitude level. Finally,  $\sigma$  is the standard deviation.  $Im$  refers to the image matrix.

### 2.5 Protein crystal classification with a neural network

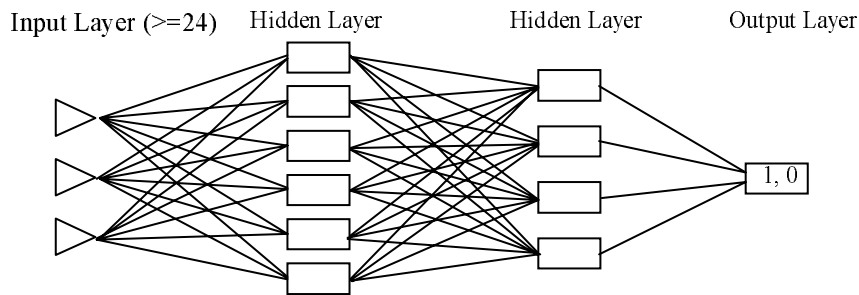


Figure 5: Three-layer Neural Network

A feed forward neural network (Figure5) was built with an input layer, two hidden layers and one output layer. Log sigmoid transfer functions were used for both the hidden layers and the output layer to generate real valued outputs between the values “0” and “1.” Training was performed with mean square error optimization and backprojection [7] [8]. Output values were grouped into two binary classes “0” for the “clear” group and “1” for “non-clear” group. A “clear” drop was defined as an image devoid of any precipitates or crystals. Examples of clear drops including empty wells, dried drops with skin formation, speckled precipitate, etc are shown in Figure 6. A “non-clear” drop was defined as a drop containing precipitates, micro-crystals or crystals and examples are shown in Figure 7.

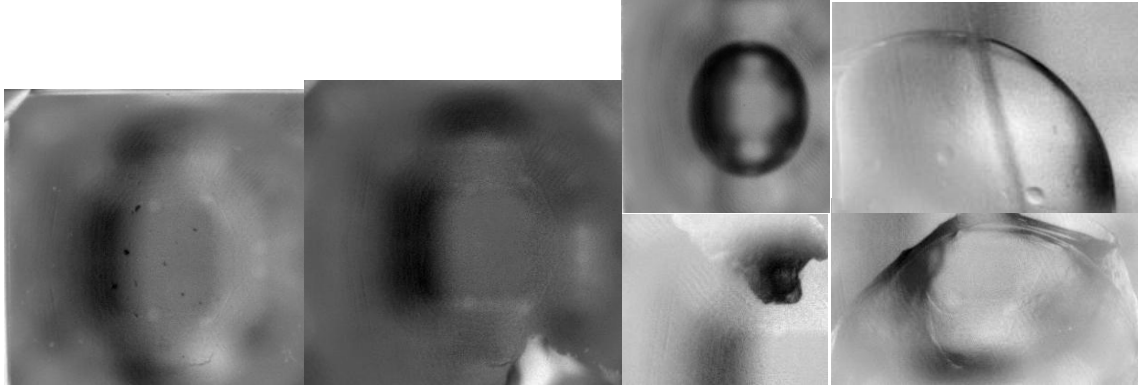


Figure 6: Examples of clear drops.

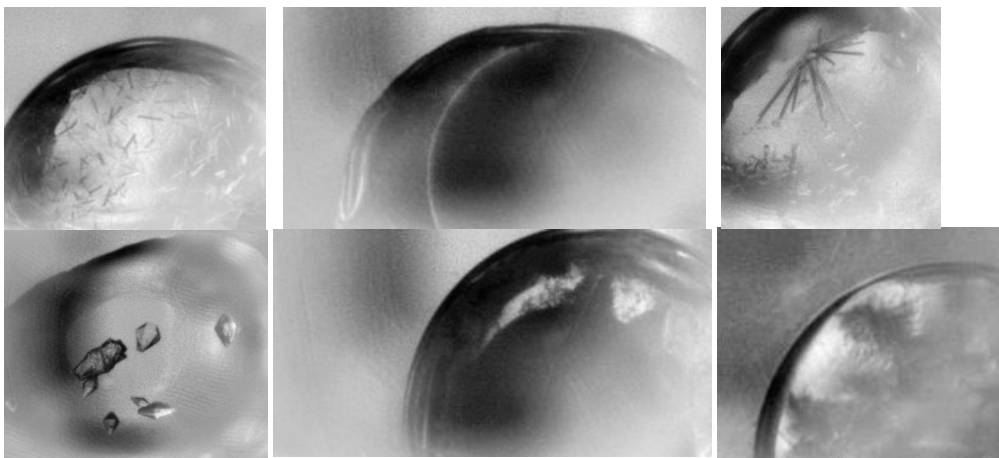


Figure 7: Examples of non-clear drops.

### 3. Experimental Results

#### 3.1 Classification

The training dataset included over 2,000 samples, and the testing dataset included over 400 samples. Classification accuracy was measured from the fraction of True Positive (TP) and True Negative (TN) rates, defined as:

- TP= True Positive (Percentage) = (correctly classified 'non-clear' images) / (all 'non-clear' images)\*100%,
- TN = True Negative (Percentage) = (correctly classified 'clear' images) / (all 'clear' testing)\*100%,

And Accuracy = (TP+TN)/2.

Using human classified images as ground truth, the current experimental results gave about 86% false positive and 94% false negative rates for classification of images with oil drops being "clear" or "non-clear".

Thus we claim an overall rate of accuracy  $(86\%+94\%)/2=90\%$ .

The misclassified images were analyzed to identify why the algorithm failed. In some cases it was even difficult for a human expert to differentiate between 'skin' and 'precipitate'. In other misclassified images, the shadow besides the boundary covered small crystals and leading to missing boundary information. For some irregular drops, the cropping did not work very well. For some particular crystals, the boundary was difficult to detect. Some of these difficult cases are shown in Figures 8 and 9.

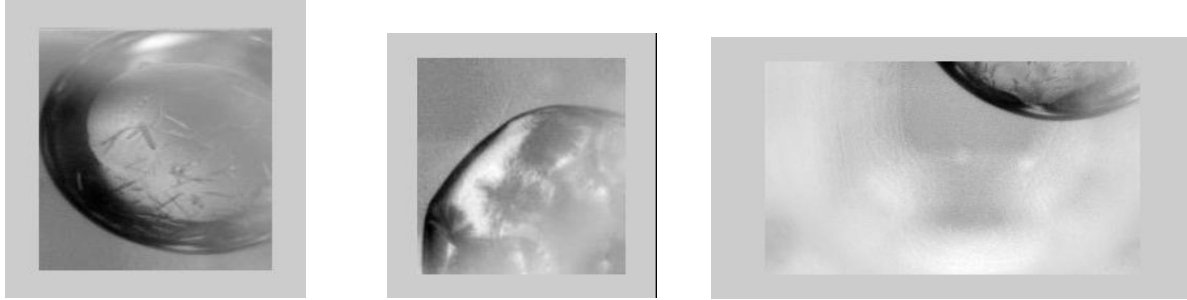


Figure 8: Non-clear drops classified as clear drops due to strong boundary shadow, lack of sharp gradient, or poor cropping.

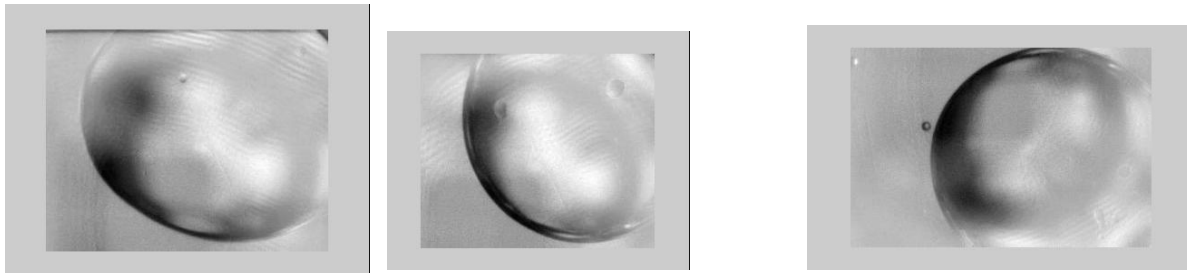


Figure 9: Clear drops classified as non-clear drops due to skin effects, boundaries influence, and boundary shadow.

### 3.2 Image database and remote Matlab® Web Server

We have established a large image protein crystal database which contains almost 10,000 micro-array images, enabling training and testing of our method with a large amount of images, against other classification methods in the research community. All images are pre-cropped and image feature vectors stored in a database. This allows our online classification server to run faster. In this more refined database, images were manually classified into three clusters: “clear”, “precipitate” and “crystal”. Currently we are testing our method using 2,000 images in the training dataset.

Figure 10 shows how we chose training and testing images from the database.

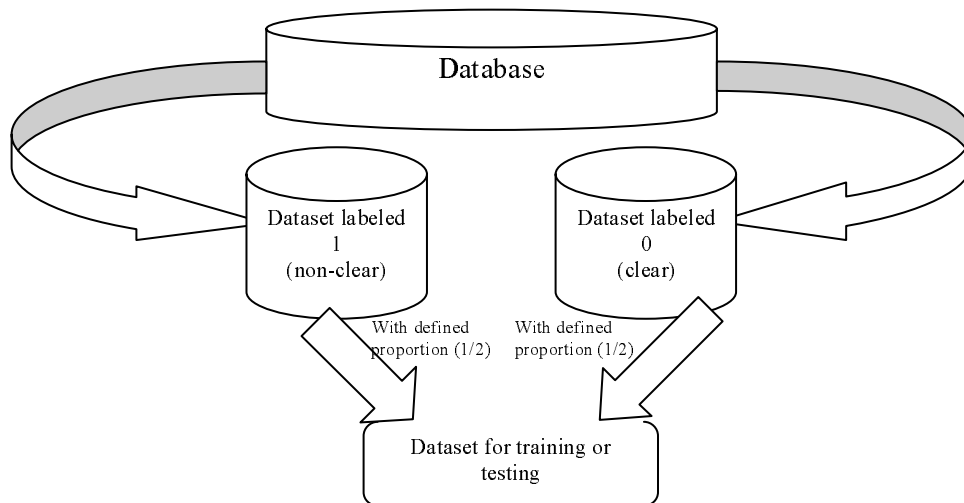


Figure 10: Setup of Genomics Image Data base for a classification experiment.

In addition, we have made substantial progress in the development of our web server for testing and training large numbers of image samples, coming from remote laboratories. A Matlab<sup>®</sup> server was installed in our imaging lab. Remote users can now use the internet to explore and provide input parameters and images through the web interface as shown in Figures 11, Figure 12 and Figure 13.

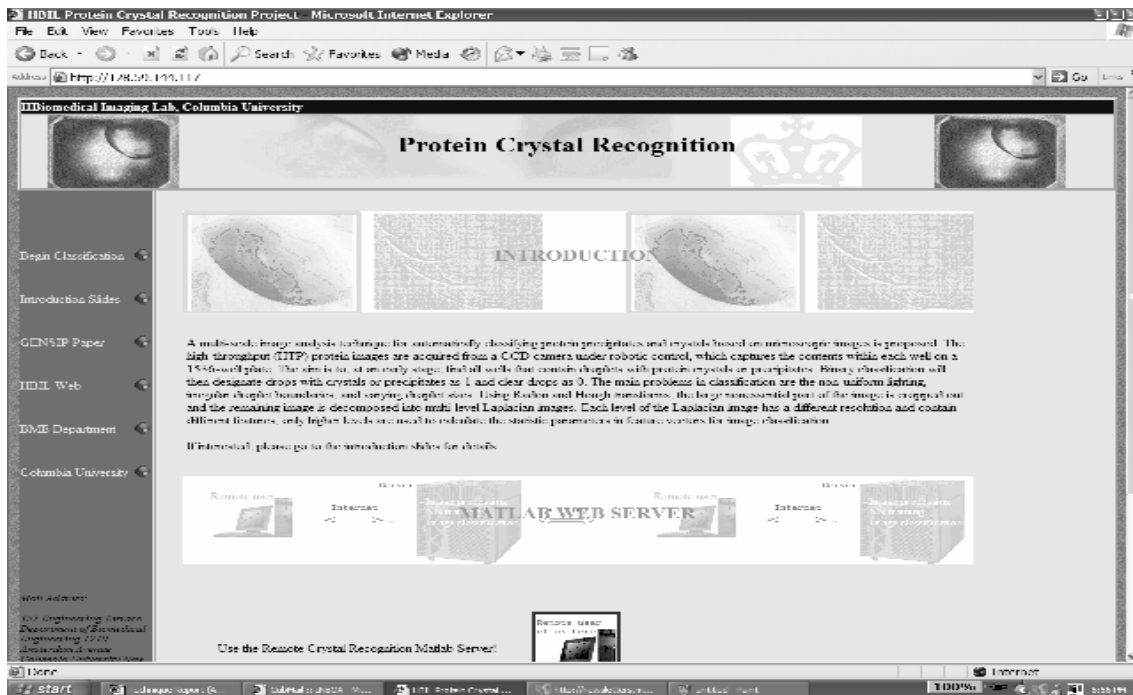


Figure 11: Web interface for classification experiment.

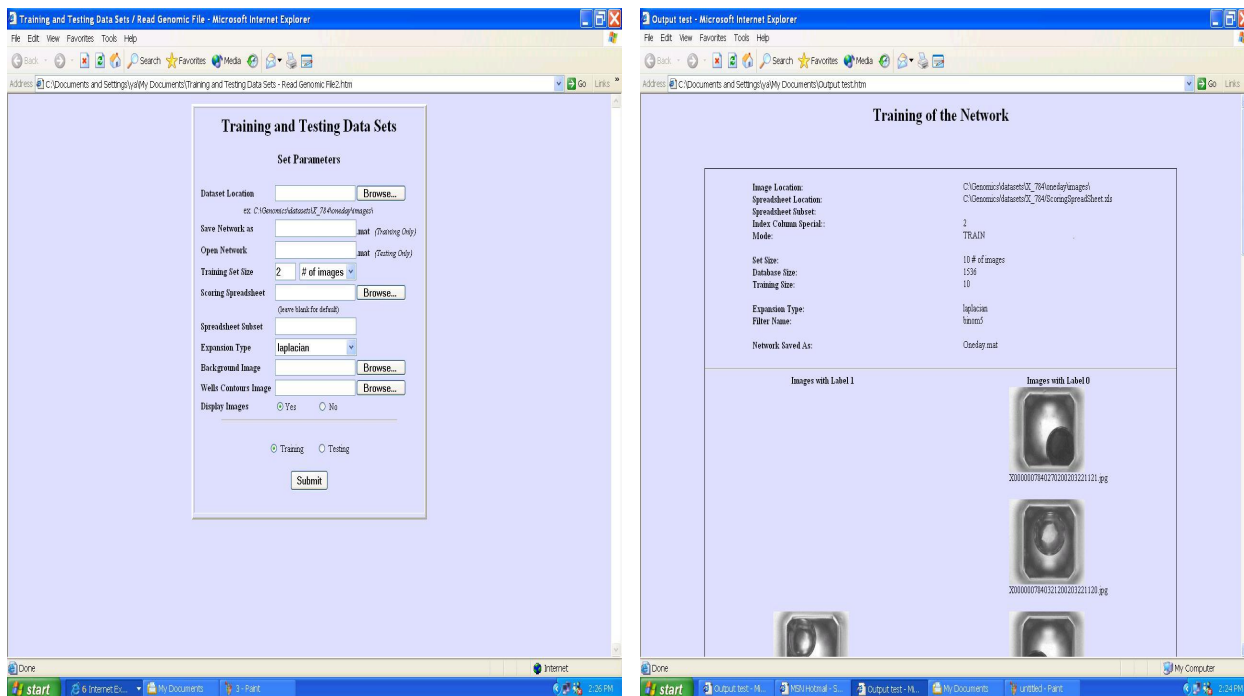


Figure 12: Crystal classification input, output interfaces.

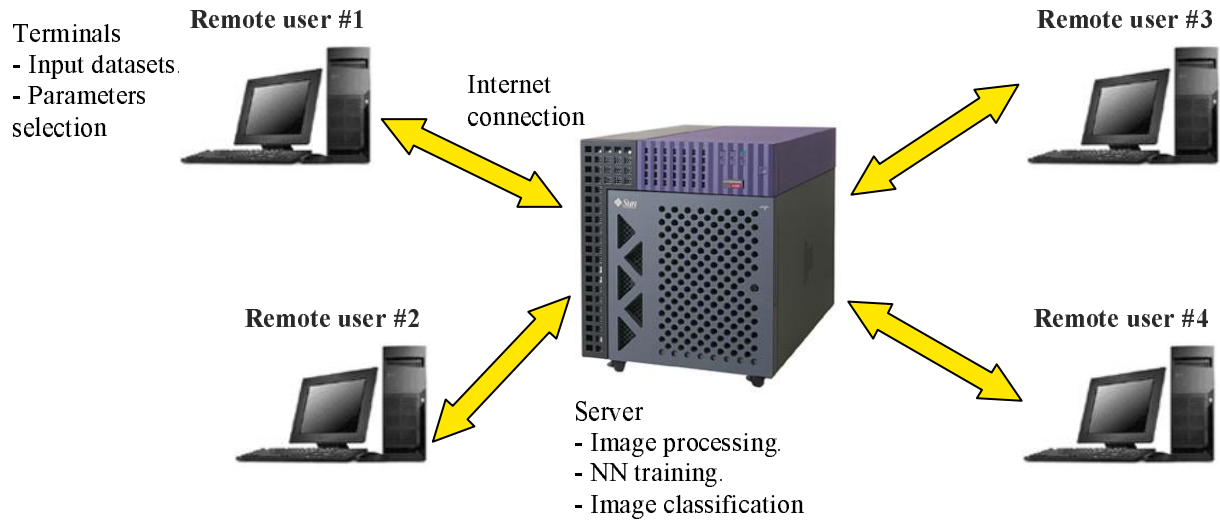


Figure 13: Matlab server for remote users.

#### 4. Conclusion

The proposed Laplacian pyramidal filters and histogram technique used in image processing and feature extraction showed promising results in classifying microscopic protein crystal images. Laplacian pyramidal filters erased redundant information from the original images and provided a representation of edge and texture patterns at distinct levels of scales. Laplacian filters were executed via a fast implementation algorithm. Use of the histogram made the features invariant to orientation which is a desirable feature since precipitates vary in forms, shapes, complexity and well orientation.

Future refinements to the classification method include the use of more training images, application of other pyramidal filters, and the further distinction of the microscopic images to separate drops with crystals, precipitates or organic matter. In addition, the creation of a web-based infrastructure for image classification facilitates collaborations between NESG institutions, the supervising research consortium, located in laboratories in the north-east corridor of the United States and Canada.

#### 5. Acknowledgement

This project is part of the Northeast Structural Genomics Consortium (NESG) sponsored by the NIH for evaluating the feasibility, costs, economics of scale, and value of an infrastructure supporting high throughput structural genomics.

#### References

1. D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, pp. 111-122, 1981.
2. P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, pp. 532-540, 1983.
3. William K. Pratt, "Digital Image Processing", Second Edition. John Wiley & Sons, Inc., c1991.
4. Tomas Hrycej, "Modular learning in neural networks : a modularized approach to neural network classification", Wiley, c1992.
5. E. D. Angelini, Y. Wang, A. F. Laine, "Classification of Micro Array Genomic Images with Laplacian Pyramidal Filters and Neural Networks", GENSIPS'04, Baltimore, MD, May 26, 2004.
6. S. Malassiotis and M. G. Strintzis, "Tracking the left ventricle in echocardiographic images by learning heart dynamics," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 282-290, 1999.
7. M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston, MA: PWS Publishing, 1996.
8. M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, pp. 525-533, 1993.