Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy

Michael V. Boland*, *Student Member, IEEE*, Mia K. Markey†, and Robert F. Murphy†

Center for Light Microscope Imaging and Biotechnology and *Biomedical Engineering Program (boland@andrew.cmu.edu) or †Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Ave. Pittsburgh, PA 15213 USA

Abstract— **We describe a method of classifying cellular protein localization patterns based on their appearance in fluorescence light microscope images. Images depicting cellular protein localization were obtained using immunofluorescence microscopy. After collection, the images were processed and subject to feature extraction. Zernike moments were calculated for each image and used as inputs to one of two classification schemes: a classification tree or a neural network. Of the two classifiers, the neural network demonstrated better performance, correctly classifying 84% of previously unseen images. This work has application as a novel approach to protein description, as a means of automating microscopes, and as part of a new approach to gene discovery.**

I. Introduction

THE goal of the work we describe here is to develop methods that allow for the numerical description and subsequent classification of the patterns found in fluorescence light microscope images of mammalian cells. Such images are generated on a regular basis by labeling one or more subcellular structures with fluorescent dyes and then collecting images of the resulting pattern of fluorescence using a microscope. The problem is then to describe these patterns in a way that is amenable to further processing by a computer.

Not all numerical descriptors (features) are equally likely to be applicable and those that are selected must meet certain criteria. The chosen features must be invariant to translations and rotations of the cell within the field of view of the microscope. They should also be insensitive to the wide variability in cell morphology that is present, even in cells subject to identical preparation. These requirements will serve to restrict investigation into appropriate features.

Much work has been done to describe and classify patterns of all sorts, but pattern recognition has been only sporadically applied to the problems of automated microscope image analysis[1], [2], [3]. One area where this is not the case can be found in the analysis of Pap smears, used in the diagnosis of cervical cancer. Pattern recognition has been extensively applied in this field, with the hope of recognizing cancerous cells in a background of normal tissue[4]. This work is fundamentally different from ours and does not provide us with a starting point. A more useful, although not biological, area to look at is that of handwriting recognition.

Recognition of handwritten characters is similar to the problem we are addressing in that although there are distinct classes of images (e.g. numbers and letters), there is considerable variability within each class (e.g. individuals' versions of the number "2"). It is therefore within character recognition that we began our search for appropriate techniques.

We envision the following applications for our methods. First, we believe that formalization of methods for quantitating protein localization will provide a new means of describing proteins. Such description will be supplementary to existing sequence analysis. Second, computational methods based on the appearance of fluorescence in the cells under study will allow for automation of microscope tasks. Finally, it will be possible to couple computational methods with molecular biology in order to automate the discovery of genes based on the localization of their proteins.

II. Materials and Methods

Collection of images was done using immunofluorescence microscopy. In short, Chinese Hamster Ovary (CHO) cells were fixed in paraformaldehyde and permeabilized with saponin before incubation with a primary antibody directed against a protein of interest. The cells were then incubated with a secondary antibody bound to a fluorescent dye. After mounting the cells on microscope slides, they were imaged onto a cooled charge-coupled device (CCD) mounted on a customized Zeiss Axiovert microscope[5]. By using four primary antibodies directed against different proteins and one DNA stain, we were able to acquire images depicting five classes of fluorescence distribution. The four proteins were giantin[6], NOP4[7], LAMP2[8], and tubulin (Sigma, St. Louis, MO USA). The DNA stain was Hoechst 33258 (Molecular Probes, Inc., Eugene, OR USA). Each field of view was acquired as a stack of three images where the focus was changed by a small amount ($0.237\mu m$) between each slice. By acquiring three dimensional data, it is possible to computationally remove any out of focus fluorescence in the middle image plane.

The images were processed by first applying numerical deconvolution to each three image stack in order to clean

up the central image plane[9]. The next step involved manually defining regions of the image that contained single cells. The background level of fluorescence, defined as the most common pixel value in the region, was then subtracted from all pixels. Fluorescent objects were identified as contiguous groups of pixels whose values were at least a constant integer multiple of the background level subtracted above. These cropped, thresholded images were then subject to feature extraction.

The features used to describe the images numerically and compactly were Zernike moments (Equation 1)[10], [11]. We chose the Zernike moments as features with the hope that they would serve as a completely general set of descriptors and would allow us to add additional image classes without redesigning our feature set. Since the Zernike polynomials are the basis set used in calculating Zernike moments, it is possible to calculate an arbitrarily large number of Zernike moments for an image. Another interesting feature of the Zernike moments is the ability to reconstruct the original image to arbitrary precision, proportional to the number of moments available. We used this aspect of the Zernike moments in order to visualize the amount and type of information that was being fed into the classifiers.

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y) V^*_{nm}(x,y) \qquad (1)$$
$$x^2 + y^2 \leq 1$$
$$n \text{ is a positive integer}$$
$$|m| \leq n \text{ and } n - |m| \text{ is even}$$
$$V^*_{nm} \text{ is the Zernike polynomial of order } n + m$$

We chose to calculate the Zernike moments up through order 12 ($m + n < 12$ in Equation 1). Since the moments themselves are complex numbers and are not insensitive to rotation of the image, we used the magnitudes of the moments as our features[11]. This provided us with 49 features that could be used as inputs to the classification schemes described below.

The first classification scheme used with the features was a classification tree, as implemented in the S-Plus (MathSoft, Seattle, WA USA) `tree()` function. This implementation is based on Classification And Regression Trees (CART), described in[12]. The second classifier was a backpropagation neural network. This classifier was implemented using PDP++[13]. The network had 49 input nodes, 20 hidden nodes, and 5 output nodes, one for each class of input.

The image feature data were separated into distinct training and test sets in order to assess the performance of the two classifiers. The classification tree was generated with the training set and its performance then measured on the test set. Training of the neural network was continued using the training data until the sum of squared error of the network on the test data was minimized.

III. Results

The performance of the two classifiers is summarized in the confusion matrices in Tables I and II. The classification tree is only modestly successful, providing only 69% correct classification on images it had not seen before. Given these results, we hypothesized that it was the limitations of the classification tree's decision boundaries that were preventing better classification. In an attempt to overcome this limitation, we implemented a backpropagation neural network. As can be seen in Table II, the performance of the neural network is significantly better than that of the classification tree, displaying an average error rate of 84%.

To gain some insight into the amount of information being captured by the first 49 Zernike moments, we chose to reconstruct images from those moments. An original image of a Hoechst stained cell and the reconstruction from its first 49 Zernike moments are shown in Figure 1. Note that there is little in the way of detailed information found in the reconstructed image. The overall shape and orientation of the original image have been captured but the texture of the nucleus has not. It should be noted that describing an image with only 49 moments represents a compression of roughly 800 to 1.

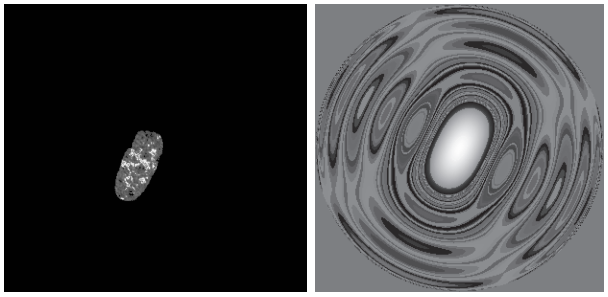

Fig. 1
Visualization of a reconstructed image. On the left is an original, thresholded image of a Hoechst stained cell. On the right is the image reconstructed from the first 49 Zernike moments of the image on the left.

IV. Discussion

The results summarized above can be considered from a pattern recognition perspective and from a biological perspective. From the standpoint of pattern recognition, we conclude that the Zernike moments represent a general method of describing the patterns found in fluorescence microscope images. They are a relatively unbiased set of features in that they are not designed to capture any "intuitive" aspects of the image data. We consider this to be advantageous in that we cannot predict with certainty

TABLE I

Performance of the Classification Tree

		Output of the Classification Tree				
		Giantin	**Hoechst**	**LAMP2**	**NOP4**	**Tubulin**
True Classification	**Giantin**	24/30 (**80%**)	1/30		3/30	2/30
	Hoechst	7/30	23/30 (**77%**)			
	LAMP2	6/60	2/60	36/60 (**60%**)	13/60	3/60
	NOP4			4/8	4/8 (**50%**)	
	Tubulin			3/26	3/26	20/26 (**77%**)

TABLE II

Performance of the Neural Network

		Output of the Neural Network				
		Giantin	**Hoechst**	**LAMP2**	**NOP4**	**Tubulin**
True Classification	**Giantin**	30/30 (**100%**)				
	Hoechst		29/30 (**97%**)		1/30	
	LAMP2	4/60		44/60 (**73%**)	9/60	3/60
	NOP4			1/8	7/8 (**88%**)	
	Tubulin			7/26	2/26	17/26 (**65%**)

the character of any additional patterns that we will add in the future.

The reconstruction exercise depicted in Figure 1 is interesting in that there is seemingly so little information present in the reconstructed image. Despite this fact, both classifiers are able to utilize this information to achieve useful levels of performance.

From a biological perspective, the performance of the classifiers is surprising given the high degree of variability present in the images. The best average performance is 84% using the neural network as the classifier. This is not as good as the classification rates obtained by investigators in other areas of pattern recognition, but the applications we envision do not require rigorous single-event classification. We have the advantage of being able to acquire many images from populations of identically prepared, biologically homogeneous cells before generating a final classification. In other words, we can generate classifications for many individual cells and then decide on a classification for the population based on a 'majority rule' determination. It is therefore possible to generate acceptable classification results even though the probability of an erroneous classification of a single cell is, in some cases, high. By considering the classification of many individual cells from a population and then pooling those classifications, it will be possible to obtain an arbitrarily high rate of classification.

There are many methods that must be developed before the applications described here can be fully realized. First of all, it will be necessary to automate the identification of individual cells in a particular field of view. The manual nature of this step will be a significant hindrance to complete automation of the process. Second, it will be advantageous to identify and test other feature sets in an effort to improve the rate of correct classification. This will allow investigators to provide fewer cells in order to make a classification decision and thereby speed up the process. Finally, it is necessary to try the analysis described here with more subcellular labels as well as with other cell types.

Of the applications mentioned in the introduction, one stands out as being the most feasible and widely applicable. The ability to describe proteins based on their subcellular localization will revolutionize the study of proteins just as quantitative description of protein sequences has already done. Until sequence analysis became widely available, researchers were required to make determinations regarding the structure and function of their new protein sequences based on the limited number of proteins they had studied previously. Now that there exist databases of all known protein sequences, it is possible to send a protein sequence to a server where its similarity to other proteins will be measured and quantified. We feel that the analysis of protein localization is in the same state as sequence analysis was in the past (i.e. all determinations are made subjectively by individual investigators). By accumulating images depicting protein localization into a database and generating features that can describe those patterns, it will be possible to quantitatively classify new proteins not only based on their sequence but also on their pattern of subcellular localization. We anticipate a time when visual examination of protein localization is every bit as unnecessary as staring at sequence information.

References

[1] M. Revenu, A. Elmoataz, C. Porquet, and H. Cardot, "An automatic system for the classification of cellular categories in cytological images," in *Intelligent Robots and Computer Vision XII: Algorithms and Techniques*, Boston, MA, USA, 1993, vol. 2055 of *Proceedings of the SPIE - The International Society for Optical Engineering*, pp. 32–43.

[2] A.I. Dow, S.A. Shafer, and A.S. Waggoner, "Morphological segmentation of multi-probe fluorescence images for immunophenotyping in melanoma tissue sections," in *Intelligent Robots and Computer Vision XII: Algorithms and Techniques*, Boston, MA, USA, 1993, vol. 2055 of *Proceedings of the SPIE - The International Society for Optical Engineering*, pp. 487–498.

[3] A.I. Dow, S.A. Shafer, J.M. Kirkwood, R.A. Mascari, and A.S. Waggoner, "Automatic multiparameter fluorescence imaging for determining lymphocyte phenotype and activation status in melanoma tissue sections," *Cytometry*, vol. 25, no. 1, pp. 71–81, 1996.

[4] K.R. Castleman and B.S. White, "Optimizing cervical specimin classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 5, pp. 451–457, 1980.

[5] D.L. Farkas, G. Baxter, R.L. DeBiasio, A. Gough, M.A. Nederlof, D. Pane, D.R. Patek, K.W. Ryan, and D.L. Taylor, "Multimode light microscopy and the dynamics of molecules, cells, and tissues," *Annu. Rev. Physiol.*, vol. 55, pp. 785–817, 1993.

[6] A.D. Linstedt and H.P. Hauri, "Giantin, a novel conserved golgi membrane protein containing a cytoplasmic domain of at least 350 kda," *Mol. Biol. Cell*, vol. 4, no. 7, pp. 679–93, 1993.

[7] C. Sun and J.L. Woolford, Jr., "The yeast nop4 gene product is an essential nucleolar protein required for pre-rrna processing and accumulation of 60s ribosomal subunits," *EMBO J.*, vol. 13, no. 13, pp. 3127–3135, 1994.

[8] B.L. Granger, S.A. Green, C.A. Gabel, C.L. Howe, I. Mellman, and A. Helenius, "Characterization and cloning of lgp110, a lysosomal membrane glycoprotein from mouse and rat cells," *Journal of Biological Chemistry*, vol. 265, no. 20, pp. 12036–43, 1990.

[9] D.A. Agard, Y. Hiraoka, P. Shaw, and J.W. Sedat, "Fluorescence microscopy in three dimensions," in *Fluorescence Microscopy of Living Cells in Culture*, D.L. Taylor and Y-L. Wang, Eds., vol. 30 of *Methods in Cell Biology*, pp. 353–377. Academic Press, Inc., San Diego, CA, 1989.

[10] R.J. Prokop and A.P. Reeves, "A survey of moment-based techniques for unoccluded object representation and recognition," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 5, pp. 438–460, 1992.

[11] A. Khotanzad and Y.H. Hong, "Rotation invariant image recognition using features selected via a systematic method," *Pattern Recognition*, vol. 23, no. 10, pp. 1089–1101, 1990.

[12] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks/Cole, Monterey, CA, 1984.

[13] R.C. O'Reilly, C.K. Dawson, and J.L. McClelland, "PDP++," 1995.