A Novel Method for Cell Phenotype Image Classification

Chao Li and Ji-feng Huang*
Department of Computer Science and Technology
Shanghai Normal University
Shanghai, P. R. China
jfhuang@shnu.edu.cn, lichao_shnu@sina.cn
*Correspondence should be addressed to Ji-feng Huang, E-mail: jfhuang@shnu.edu.cn

Abstract— **As the development of human genomic project, the life science research has entered the post-genome era. The study of the function of the encoded proteins is one of the hotspots in life-science research and protein subcellular localization is an important basis for functional study of the protein. The most common method used for determining subcellular localization of protein in cell is fluorescence microscopy. Image feature calculation has proven invaluable in the automated cell phenotype image classification. This article proposes a novel method for cell phenotype image classification which is to count the local difference features of the fluorescence images. The novel method is tested on two image sets called LOCATE Endogenous and LOCATE Transfected. A support vector machine was trained and tested for each image set and better classification accuracies were obtained on the two image sets.**

Keywords-local difference features, protein subcellular location, classification, support vector machine

I. INTRODUCTION

Important and basis information for understanding the functions of tens of thousands of proteins at the cellular level is the location and distribution of protein. High-throughput automated fluorescent microscope imaging technologies provide a powerful way of acquiring such information. As bioimage data is increasingly used to understand protein function at the cellular level, a large number of images are needed. Images of protein locations are generally analyzed in traditional ways, which is time consuming and prone to errors [1, 2]. Over the past decade, however, machine learning methods have using to automate the distribution of subcellular location from fluorescence imagery instead of the traditional approach [3]. To deal with the large scale of data, we must have a rapid and effective method.

Image feature calculation has proven very useful in the automated analysis of subcellular images [4]. In combination with machine learning methods, image feature calculation has shown highly successful at analyzing and distinguishing subcellular images and has exceeded human classification accuracy. The typical machine learning methods such as neural network and support vector machines have significant performance in this field. Murphy et al. have proposed Subcellular Location Feature sets and tested the performance of these features by trained a back propagation neural network (BPNN) [5]. Boland and Murphy have trained the BPNNs to test the performance of Haralick textures, Zernick moments and SLFl [6]. Murphy et al. have employed multiple feature sets to train BPNN including of Haralick textures, Zernike moments, and morphological features [7]. Chen et al. have developed an automated method that selects the best feature set for protein subcellular localization [8]. Hamilton et al. have developed Threshold Adjacency Statistics (TASs) and trained SVM to test the performance of the features [9]. Nanni et al. have developed optimized sets of various feature extraction strategies including Local Ternary Pattern, Wavelet features, Haralick textures, TAS, and LBPs for training an ensemble using random subspace of Levenberg-Marquqrdt neural networks [10]. Qian Xu et al. have proposed multitask learning for protein subcellular location prediction [11].

However, a difficulty with these approaches is that each cell type has the diversification of organelle structure, they needed high computational cost. Another difficulty is that what features should be extracted and how to extract features form large number of the subcellular images. In location pattern recognition, cells change so greatly in their size, intensity, shape, orientationation and position that raw pixel intensity values are not very useful [8]. Invariant features were needed in image recognition systems. The purpose of this paper is to propose a novel method for feature extraction which was performed better in prediction of protein subcellular location. In image processing or pattern recognition problem, it is important to extract invariant features from given images. If the categorical property of an image would not be changed along with some transformation, the feature should be invariant to rotation and invariant to translation. In this paper we advanced a new method to extract features from the image, and these features are invariant under rotation and translation. It is a computationally simple and fast geometrical measure for discriminating protein subcellular localization. Experiments show that this new method performed excellent in protein subcellular localization.

II. MATERIALS AND METHODS

A. Image Datasets

Two image sets called LOCATE Endogenous and LOCATE Transfected were established for subcellular organelles [4, 12]. LOCATE Endogenous dataset is consist of 502 images and these images are distributed in 10 classes, and LOCATE Transfected dataset contains 553 images and they are divided into 11 classes. Each image is 8 bit grayscale and of size 768 x 512. Sample images are shown in

Figure 1 and 2. The complete image set is available for download from http://locate.imb.uq.edu.au..

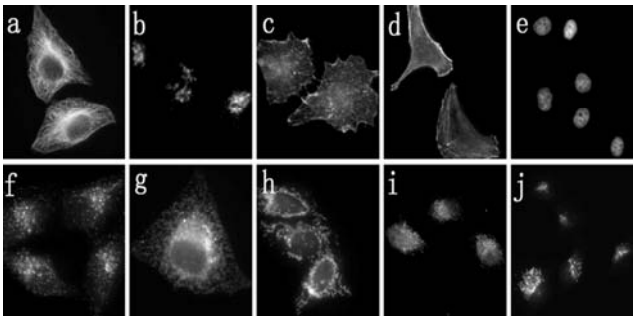

Figure 1. Sample images of the 10 organelles in LOCATE endogenous dataset.(a)Microtubule,(b)Golgi,(c)Plasmamembrane,(d)Actincytoskeleton, (e)Nucleus,(f)Endosome,(g)ER,(h)Mitochondria,(i)Peroxisome,(j)Lysosome.Scale bar 10 um.

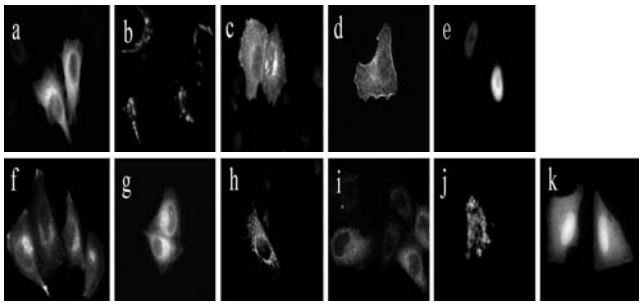

Figure 2. Sample images of the 11 organelles in LOCATE transfected dataset.(a)Microtubule,(b)Golgi,(c)Plasmamembrane,(d)Actincytoskeleton, (e)Nucleus,(f)Endosome,(g)ER,(h)Mitochondria,(i)Peroxisome,(j)Lysosome, (k) Cytoplasm. Scale bar 10 um

B. Feature Extraction Strategies

This novel method we called LDP (local difference pattern) which is based on local difference features. The features which are invariant to rotation and invariant to translation of the pixels in the image here were derived from morphological and geometric image analysis. Local difference features, which were calculated the differences between the gray value of the central pixel c and the gray values of P pixels in the neighborhood, were generated by first applying a threshold to the image to divide into two components, one is background, and another is object. P is defined as 8 in this work. But it is different form binary image. To all pixels in the image, the grey value which is greater than the threshold value was remained unchanged; the other which is less was assigned to 0. The threshold was chosen as follows. The mean value, μ, of those pixels with grey value at least θ is calculated for the image. And threshold is determined which is equal to μ-θ. μ is generated according to Equation 1.

$$\mu = \frac{\sum_{n=1}^{N} x_n}{N}, x_n \geq \theta \qquad (1)$$

x_n shows the value of the pixel which is greater than θ in the image. N represents the number of pixels whose values are greater than θ in the image. θ is the user defined threshold, and in this work θ is equal to 30.

We can see the grey value of pixels in the background component of the image is 0 and in the objective component of the image is from u-30 to 255 (Figure 3a' and 3b'). The range was selected to maximize the visual difference of preprocessed images for which the images had different localization but were visually similar (Figure 3). Then thirty-six statistics were obtained from the image in total. The thirty-six statistics were designed as follows, for each pixel in the image, whose values are non-zero, the difference between its grey values and surrounding eight values was calculated. Then we marked 1 to these surrounding pixels whose value is greater than the middle's and 0 is to those whose value is less(Figure 4 (0)-(8)). The number of surrounding pixels marked with 1 is counted for each pixel whose value is non-zero. Next, the first statistic is then the number of pixels with no neighbor marked with 1; the second is the number with one neighbor marked with 1, and so forth up to the maximum of eight, nine statistics are obtained and normalized by dividing each by the total number of pixels whose values are non-zero in the threshold image. To each pixel in the image, we calculate the D-value of the eight grey values with the central one comparatively and sum up the values above 0 and those below 0, and the pixels have a same type are classified as a category as shown in Figure 4, so each image has nine statistics of values above 0 and nine statistics of values below 0 and eighteen statistics were obtained here. Then to these pixels of each category as shown in Figure 3, we sum up the eight D-value (absolute value) for every pixel, and nine variances are calculated on the nine categories, so nine statistics are generated for the image. Finally, thirty-six features are obtained for the image.

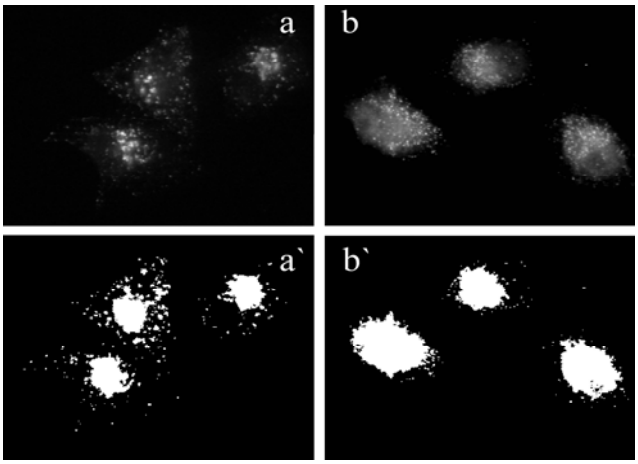

Figure 3. Distinguishing cell images by threshold. Images (a) and (b) are texturally and visually similar, but images (a') and (b') are more distinguished. Image (b`) contains more solid white regions, while (a`) shows more external speckling and feathering of edges.

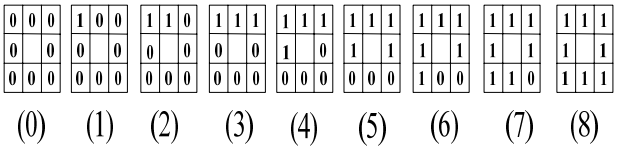

Figure 4. 8-Neighbor statistics for cell images.The first statistics is then the number of pixels with zero neighbors marked with 1, the second is the number with one neighbor marked with 1, and so on up to eight.

III. Results and Discussion

A. Results

The efficacy of new method called LDPs in predicting subcellular localization was then tested by generating statistics for the endogenous and transfected images, and creating a SVM for each. Libsvm software was using to create SVMS [13]. The endogenous images set have ten kinds of organelles including 503 images; while the transfected images set has eleven kinds of organelles including 545 images. The data set was randomly split into two sections, one is for training and the other is for testing. Each data set was split into 4/5 for training and 1/5 for testing randomly. A SVM was then trained on the training set and by localization class classification accuracies on the testing were recorded. Random data splitting training and testing was then repeated 1000 times. The overall average classification accuracy on endogenous test sets and transfected test sets were then 96.7% and 92.3% (Table 1), respectively.

TABLE I. Comparison of Haralick, TAS, LDP statistics classification accuracies

Data Set	Haralick	TAS	LDP
Endogenous	94.2%	94.4%	96.7%
Transfected	86.0%	86.6%	92.3%

B. Discussion

The novel approach is simple, accurate, and effective for protein subcellular location images from the two LOCATE datasets. It has been shown that the performance of classification is better compared to *TAS* and *Haralick texture* [9]. We have utilized both translation and rotation invariant features, and these features are different from each class obviously.

IV. Conclusions

The function of the encoded proteins is attracting ones' attention increasingly. Subcellular localization can provide useful information for improving predictions of protein conformation. While image statistics have proved highly successful in distinguishing, here we propose a new method based on invariant of translation and rotation for feature extraction which is calculating the D-value of the eight pixels' grey values with the central one comparatively. They remove the need for cropping of individual cells from images and with a classification up to 97% and 92.3% they offer better accuracy than *TAS* and *Haralick texture*[5, 9], while having a fast speed to calculate, both basic requirements for application to large-scale approaches.

Acknowledgment

The work was supported by the Program of Shanghai Normal University (DZL126).

References

[1] A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, R. F. Murphy and J. Kovacevic, "A multiresolution approach to automated classification of protein subcellular location images," BMC Bioinformatics, 2007, 8:210.

[2] L. Nanni, A. Lumini, Y. –S. Lin, C. –N. Hsu, and C. –C. Lin, "Fusion of systems for automated cell phenotype imae classification," Expert Systems with Applications, 2010, 37, pp. 1556-1562.

[3] E. Glory and R. F. Murphy, "Automated subcellular location determination and high-throughput microscopy," Development Cell, 2007, January, pp. 7-16.

[4] N. A. Hamilton, J. T. Wang, M. C. Kerr and R. D. Teasdale, "Statistical and visual differentiation of subcellular imaging," BMC Bioinformatics, 2009, 10:94.

[5] R. F. Murphy, M. V. Boland and M. Velliste, "Towards a systematics for protein subcellular location: Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images," In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 2000, AAAI Press, pp. 251-259.

[6] M. V. Boland and R. F. Murphy, "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells," Bioinformatics, 2001, 17, pp. 1213-1223.

[7] R. F. Murphy, M. Velliste and G. Porreca, "Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images," Journal of VLSI Signal Processing, 2003, 35, pp. 311-321.

[8] X. Chen and R. F. Murphy, "Objective clustering of proteins based on subcellular location patterns," J Biomed Biotech, 2005, 2, pp. 87-95.

[9] N. A. Hamilton, R. S. Pantelic, K. Hanson and R. D. Teasdale, "Fast automated cell phenotype image classification," BMC Bioninformatics, 8.

[10] L. Nanni, S. Brahnam and A. Lumini, "Novel features for automated cell phenotype image classification," Advances in Computational Biology: Advances in Experimental Medicine and Biology (AEMB), 2010, 680, pp. 207-213.

[11] Q. Xu, SJ. Pan, HH. Xue and Q. Yang, "Multitask learning for protein subcellular location prediction," Transaction on Computational Biology and Bioinformatics, 2011, Vil.8, pp. 748-758.

[12] J. L. Fink, R. N. Aturaliya, M. J. Davis, F. Zhang, K. Hanson, M. S. Teasdale and R. D. Teasdale, "LOCATE: A protein subcellular localization database," Nuci Acids Res 2006, 34((database issue)).

[13] LIBSVM: a library for support vector machines. http://www.csite.ntu.edu.tw/~cjlin/libsvm.