# Subcellular Localization Algorithm Based On Fluorescence Microscopy Images

Chao Li, Xue-Hong Wang, Li Zheng, Ji-Feng Huang\*

Department of Computer Science and Technology Shanghai Normal University \*Corresponding author: jfhuang@shnu.edu.cn

**Abstract.** To understand the function of the encoded proteins, we need to be able to know the subcellular location of a protein. The most common method used for determining subcellular location is fluorescence microscopy which allows subcellular localizations to be imaged in high throughput. Image feature calculation has proven invaluable in the automated analysis of cellular images. This article propose a novel method for feature extraction from given images, the nature which is to count the number of above threshold pixels, the threshold is given by calculating the average intensity of those pixels with intensity at least 30,and calculate the D-value(difference value) of the eight pixels' grey values with the central one comparatively. They are tested on images set (available download) which fluorescently tagged proteins are endogenously expressed in 10 subcellular locations and classification accuracies of 96.7% are obtained on the endogenous set.

## 1 Introduction

To understand the functions of tens of thousands of proteins at the cellular level, we need to obtain data about the subcellular distributions of proteins and subsequent identification of the encoded proteome. High-throughput automated fluorescent microscope imaging technologies provides a powerful way of acquiring such information. As bio image data was increasingly used to understand protein function at the cellular level, vast numbers of images including multiple fluorophores for cells under a variety of experimental conditions. It is estimated that having a single image for every combination of cell type, protein, and timescale would require on the order of 1000 billion images [1]. Traditionally, the primary means of analyzing and annotating images depicting subcellular location in these large-scale studies has been visual examination [2]. Over the past decade, however, the traditional approach has begun to be replaced with using machine learning methods to automate the determination of subcellular location from fluorescence microscope images. To deal with the scale of the data becoming available automated annotation, analysis, comparison, classification and storage of cellular images is essential.

Image feature calculation has proven invaluable in the automated analysis of cellular images [3]. Measures of features such as texture and morphology may be used to generate a vector of numbers for a given cell image, and in combination with machine learning methods such as neural network and support vector machines they have proved highly successful at classifying subcellular images of the major organelles of a cell and have exceeded human classification accuracy.

While image feature calculation has performed well in subcellular localization classification, a difficulty with these approaches is that organelle structure can vary widely between each cell type, hence 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 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 distinguishing subcellular localization. Experiments show that this new method performed excellent in subcellular localization.

### 2 Results

#### 2.1 Image Datasets

An image set called Endogenous was established for subcellular organelles. The set for which an endogenous protein or feature of the specific organelle was detected with a fluorescent antibody or other probe (10 organelles).Each image in the set was accompanied by an additional image of the cells counterstained with the DNA specific dye 4', 6-diamidino-2-phenylindole (DAPI), which highlights the location of the nucleus of every cell in the image. In addition, the DAPI image was reviewed to exclude images that contained one or more cells not in interphase. Each organelle set consists of about 50DAPI counterstained images. In total 502 endogenous localization images were obtained. All images were of fixed HeLa cells, and taken at  $60 \times$  magnification under oil immersion. They are 8 bit grayscale, 768 by 512 pixels, each containing up to 13 cells. Sample images of 10 organelles are shown in Figure 1.The complete image set is available for download from http://locate.imb.up.edu.au[4, 5].



**Figure.1.** Sample images of the 10 organelles.(a)Microtubule,(b)Golgi,(c)Plasma membrane,(d)Actincytoskeleton,(e)Nucleus,(f)Endosome,(g)ER,(h)Mitochondria,(i)Peroxisome,(j)L ysosome.Scale bar 10 um.

#### 2.2 Features for Classification

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 [6, 7]. 8-Neighbor statistics, which were calculated the D-value of the eight grey values with the central one comparatively for each pixel in the image, were generated by first applying a threshold to the image to divide into two components, one is background, and another is object. 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, u, of those pixels with grey value at least 30 is calculated for the image. And threshold is determined which is equal to u-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 2a' and 2b'). The range was selected to maximize the visual difference of preprocessed images for which the images had different localization but were visually similar (Figure 2). 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 3 (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 3, 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.2.** Distinguishing cell images by threshold. Images of the endosome (a) and the peroxisome (b) are preprocessed by threshold (a' and b') so that pixels with grey value in the range u-30 to 255 are shown in white, where u is the mean pixel value of each image. 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.3.** 8-Neighbor statistics for cell images. To these pixels, which values is non-zero, the number of the surrounding neighbors which are marked with 1 is counted. Examples of having zero to eight neighbors which marked with 1 are given in (0)-(8). 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.

#### 2.3 Classification with the Features

The efficacy of new features in predicting subcellular localization was then tested by generating statistics for the endogenous images, and creating a SVM for it. Libsvm software was using to create SVMS [8, 9]. The endogenous images set have ten kinds of organelles including 503 images. The thirty-six statistics of ten sample images, which were belong to one of the ten organelles, were shown in Figure 4. There are different from each other we can find from Figure 4. The data set was randomly split into two sections, one is for training and the other is for testing. The training set contained 400 images. And the testing one include of 103 images. 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 the 1000 endogenous test sets was then 96.7%. The classification accuracies for each class of localization for the endogenous set were all high being in the range 91.1% to 100% (Table 1). The class with lower classification accuracy such as Golgi, Mitochondria and Lysosome appear to be those that exhibit higher visual similarity to each other than to other classes.



**Figure.4.** Thirty-six statistics of ten sample images which were belong to one of ten organelles. Our method to distinguish and classify the subcellular images was based on the D-value of the pixels in the image. And the thirty-six statistics are shown.

| Table 1.Average classification accuracies using new method on Endogenous data test sets. The       |
|--|
| numbers in the table are representing the probability of the test data assigned to the correspond- |
| ing class (repeated 1000 times).   |

| True<br>Classification | Output of the Classifier |       |       |          |              |         |       |                        |             |            |
|------------------------|--------------------------|-------|-------|----------|--------------|---------|-------|------------------------|-------------|------------|
|                        | Endosome                 | ER    | Golgi | Lysosome | Mitochondria | Nucleus | PM    | Actin-<br>Cytoskeleton | Microtubule | Peroxisome |
| Endosome               | 98.6%                    | 0     | 0.4%  | 0.8%     | 0            | 0       | 0.2%  | 0                      | 0           | 0          |
| ER                     | 0                        | 94.7% | 2.7%  | 0        | 0            | 0       | 2.6%  | 0                      | 0           | 0          |
| Golgi                  | 0                        | 2.9%  | 91.6% | 3.4%     | 2%           | 0       | 0.1%  | 0                      | 0           | 0          |
| Lysosome               | 0                        | 0     | 0.9%  | 91.1%    | 5.9%         | 2.1%    | 0     | 0                      | 0           | 0          |
| Mitochondria           | 0                        | 0     | 0     | 5.2%     | 94.8%        | 0       | 0     | 0                      | 0           | 0          |
| Nucleus                | 0                        | 0     | 0     | 0        | 0            | 99.8%   | 0     | 0.2%                   | 0           | 0          |
| РМ                     | 1.6%                     | 1.6%  | 0     | 0        | 0            | 0       | 94.8% | 1.8%                   | 0           | 0.2%       |
| Actin-<br>Cytoskeleton | 0                        | 0     | 0     | 0        | 0            | 0       | 0     | 100%                   | 0           | 0          |
| Microtubule            | 0                        | 0     | 0     | 0        | 0            | 0       | 0     | 0                      | 100%        | 0          |
| Peroxisome             | 0                        | 0     | 0     | 0        | 0            | 0       | 0     | 0                      | 0           | 100%       |

## **3** 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 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% they offer better accuracy than *TAS* and *Haralick* [4,10], while having a fast speed to calculate, both basic requirements for application to large-scale approaches.

### References

 L. P. Coelho, E. Glory, J. Kangas, S. Quinn, A. Shariff and R. F. Murphy.: Principles of Bioimage Informatics: Focus on Machine Learning of Cell Patterns. Lecture Notes in Computer Science, Vol.6004. Springer-Verlag Berlin Heidelberg (2010) 8-18

- E. Glory and R. F. Murphy.: Automated Subcellular Location Determination and High-Throughput Microscopy. Developmental Cell. January (2007) 7-16
- 3. N. A. Hamilton, J. TH. Wang, M. C. Kerr and R. D. Teasdale.: Statistical and visual differentiation of subcellular imaging. BMC Bioinformatics. 10:94(2009)
- 4. N. A. Hamilton, R. S. Pantelic, K. Hanson and R. D. Teasdale.: Fast automated cell phenotype image classification. BMC Bioinformatics. 8:110(2007)
- J. L. Fink, R. N. Aturaliya, M. J. Davis, F. S. Zhang, K. Hanson, M. S. Teasdale, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki and R. D. Teasdale.: LOCATE: a mouse protein sub-cellular localization database. Nucleic Acids Research, Vol. 34(2006) 213-217
- T. Kurita, N. Otsu and T. Sato.: A Face Recognition Method Using Higher Order Local Autocorrelation and Multivariate Analysis. PATTERN RECOGNITION, Vol. II. Conference B. The Hague (1992) 213-216
- X. Chen and R. F. Murphy.: Objective Clustering of Proteins Based on Subcellular Location Patterns. Journal of Biomedicine and Biotechnology.2(2005)87-95
- 8. LIBSVM: a library for support vector machines. http://www.csite.ntu.edu.tw/~cjlin/libsvm
- K. Huang and R. F. Murphy.: Automated classification of subcellular patterns in multicell images without segmentation into single cells. Proceedings of the IEEE 2004. (2004)1139-1142
- R. M. Haralick.: Statistical and structural approached to texture. Proceedings of the IEEE 1979. 67(1979)768-804