AUTOMATED CLASSIFICATION OF SUBCELLULAR PATTERNS IN MULTICELL IMAGES WITHOUT SEGMENTATION INTO SINGLE CELLS

Kai Huang[1,3] and Robert F. Murphy[1,2,3]

Departments of Biological Sciences[1] and Biomedical Engineering[2], and Center for Automated Learning and Discovery[3], Carnegie Mellon University, 4400 Fifth Ave., Pittsburgh, PA, 15213, USA, murphy@cmu.edu

ABSTRACT

Fluorescence microscope images capture information from an entire field of view, which often comprises several cells scattered on the slide. We have previously trained classifiers to accurately predict subcellular location patterns by using numerical features calculated from manually cropped 2D single-cell images. We describe here results on directly classifying fields of fluorescence microscope images using a subset of our previous features that do not require segmentation into single cells. Feature selection was conducted by stepwise discriminant analysis (SDA) to select the most discriminative features from the feature set. Better classification performance was achieved on multicell images than single-cell images, suggesting a promising future for classifying subcellular patterns in tissue images.

1. INTRODUCTION

Subcellular location is an important characteristic of proteins in that it represents the biochemical context of protein functionality. With the development of highly sensitive probes and high-throughput imaging instruments, fluorescence microscopy has become the major tool to study protein subcellular distributions. Until recently, visual inspection was the only way to distinguish different subcellular patterns in microscope images. Our group has developed systematic approaches to describe subcellular location, including building classifiers that can recognize all major subcellular patterns in single-cell fluorescence microscope images [1-4]. Our single-cell image features cover a variety of subcellular distribution properties such as morphology, texture, moments, and geometry. About 92% and 96% average accuracy over ten major subcellular structures were achieved for 2D and 3D single-cell images respectively (Huang and Murphy, submitted). Although several automatic cell segmentation methods have been described [5-7], they either require extra labeling of the nucleus and plasma membrane or depend on specific model assumptions that do not perform very well for arbitrary fluorescence microscope images. Instead, we would like to classify protein subcellular distribution patterns directly on multicell images, each of which contains a single protein labeled in different cells.

2. MULTICELL IMAGE CLASSIFICATION

2.1 Multicell Image Features

To describe the subcellular distribution of a protein in multicell images, robust features are needed that are independent of the number and rotation of cells. As an initial approach, we identified the features from our most powerful current feature set, SLF7 [8], that do not require segmentation into single cells.

Morphological Features

13 single-cell morphological features (Table 1) that are independent of the number of cells were selected from SLF7. These features describe the property of objects and edges in an image and therefore are independent of cell rotation as well.

Haralick Texture Features

The 13 Haralick texture features used for single-cell image classification [1] all apply equally well to multicell images in that they represent statistics of co-occurrences of gray-level pixels. In calculating Haralick texture features, four gray-level co-occurrence matrices from horizontal, vertical, and two diagonal directions are averaged to achieve rotation invariance. Since texture essentially represents repetitive local patterns in an image, texture features are invariant of the number of cells and should even work for partial cells. Haralick texture features are affected by the number of gray levels and the pixel size in an image. We have previously shown that the most discriminative Haralick texture features for classifying subcellular patterns can be generated using 256 gray levels for an image that has a pixel width of 1.15 micron [8].

0-7803-8388-5/04/$20.00

SLF Index	Multicell Morphological Feature Description
SLF1.3	The average number of pixels per object
SLF1.4	The variance of the number of pixels per object
SLF1.5	The ratio of the size of the largest object to the smallest
SLF7.9	The fraction of the non-zero pixels in a cell that are along an edge
SLF7.10	The fraction of all values that fall in the first two bins of the edge intensity histogram
SLF7.11	The ratio of the largest to smallest value in the edge intensity histogram
SLF7.12	The ratio of the largest to the next largest value in the edge intensity histogram
SLF7.13	The edge direction difference
SLF7.80	The average length of the morphological skeleton of objects
SLF7.81	The average ratio of object skeleton length to the convex hull area of the skeleton
SLF7.82	The fraction of object pixels contained within the skeleton
SLF7.83	The fraction of object fluorescence contained within the skeleton
SLF7.84	The ratio of the number of branch points in the skeleton to the length of the skeleton

Table 1. Descriptions of multicell morphological features.

Since most modern fluorescence microscopes can acquire images with resolution and pixel intensity higher than these values, we suggested that for analysis of subcellular patterns that input images should be re-sampled to 1.15 micron/pixel and re-quantized in 256 gray bins before Haralick texture feature calculation. This procedure was therefore followed here.

2.2 Support Vector Machine Classifier

Support vector machines (SVM) represent a generalized linear classifier that looks for the maximum margin hyperplane in the feature space after transformation by a kernel function [9]. Given a learning problem described in a feature space, the task of a classifier is to find the optimal decision boundary that minimizes some predefined cost function. More often than not, the decision boundary is nonlinear and hard to represent in closed form. SVM employs a kernel function, which is often nonlinear, to map the dataset from the original feature space to a very high, sometimes unlimited, dimensional space, where the previous decision boundary becomes linear. The choice of the maximum margin hyperplane in the transformed feature space guarantees minimization of the upper bound of the expected prediction error of the SVM. Compared to other classifiers such as neural networks and AdaBoost, SVM is faster to train and therefore is often used in tasks that require heavy parameter tuning.

Several methods have been described to extend the original binary SVM [9-11]. The max-win strategy employs k binary SVMs in a k-class problem, each of which separates class j from non-j. The class with the highest value is chosen. The pair-wise strategy trains k(k-1)/2 binary SVMs for each class pair. The class voted for most frequently by all binary SVMs is selected as the output. The DAG method instead put the above k(k-1)/2 binary SVMs in a rooted binary DAG. The loser class gets removed at each node and the surviving class after tracing down from the root to a leaf node is selected as the output. We evaluated SVMs of Gaussian kernel with different variances and error penalties under all three multi-class strategies. The evaluation was conducted by using an SVM toolbox downloaded from http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox/ along with 10-fold cross validation.

2.3 Feature Selection

The features described above might contain redundancies. Some of them might also be non-discriminative (i.e., not able to contribute to the classification task). Our previous results have shown that feature reduction can increase classification accuracy as well as speed up a classifier [1, 3]. Among a group of eight different feature reduction methods, stepwise discriminant analysis (SDA) was ranked as the best feature selection method for subcellular pattern classification [3]. We therefore applied SDA to improve our classification performance by selecting the most discriminative features from our feature set.

2.4 Image Dataset

To test the feasibility of directly classifying subcellular patterns in multicell images, we created a set of images with various numbers of cells from cropped single HeLa cell images. These include fluorescence microscope images representing ten major subcellular structures [2]. Each cropped image was preprocessed by subtracting image background followed by generating a rectangular bounding box surrounding the cell region. 2-6 these bounding boxes were mixed randomly to form a multicell image. 50 multicell images were created for each class, resulting in a total of 500 multicell images which have various numbers of cells (each multicell image contained

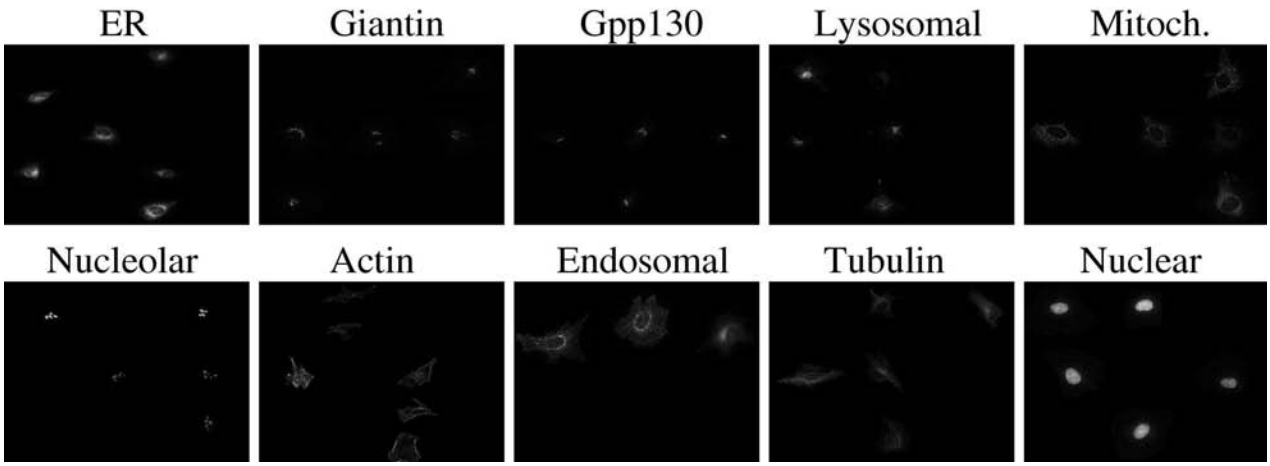

Figure 1. Example multicell HeLa images depicting ten major subcellular location patterns. The targets labeled include an endoplasmic reticulum protein, two Golgi proteins Giantin and Gpp130, LAMP2 in lysosomes, a mitochondria outer membrane protein, the nucleolar protein nucleolin, f-actin and tubulin in the cell skeleton, transferring receptor from endosomes, and DNA.

only one of the ten subcellular patterns). Figure 1 displays an example image for each pattern.

3. EXPERIMENTAL RESULTS

3.1 Classification Results

Each of the two feature subsets described above and their combination were first evaluated using a DAG Gaussian-kernel SVM with a variance level 0.01 and error penalty 20 [3] along with 10-fold cross validation. The results are summarized in Table 2. The Haralick texture feature subset achieved much higher precision and recall than the morphological feature subset. The combination of Haralick and morphological features yielded the highest precision and recall among the three feature sets. Less than 5% improvements were observed for the combined feature set over the Haralick feature subset alone, suggesting that redundancy existed between the two feature subsets.

To find the optimal parameter setting for a Gaussian-kernel SVM, we employed the entire feature set composed of both the Haralick and morphological features to evaluate Gaussian-kernel SVMs with 10-fold cross validation across many different variance levels and error penalties under all three multi-class strategies. None of the pair-wise multi-class strategy gave an average recall over 20%, suggesting that the assumption of this approach, that is no bias is given to an unknown class for each binary classifier, does not hold in our problem. The max-win and DAG strategies achieved close performance on most parameter choices. The resulting recalls showed that the optimal parameters for a Gaussian-kernel SVM on our multicell image dataset with the entire feature set were composed of a variance level of 0.01, an error penalty of 25, and the max-win multi-class strategy. A marginal improvement in performance over that in Table 2 was achieved using the parameter-tuned SVM, which yielded a precision and recall of 94.8% and 94.8% respectively. The confusion matrix is displayed in Table 3, where the two Golgi proteins, giantin and gpp130 (which we have shown are visually indistinguishable [8]), can be distinguished with a recall over 98%. The endosomal and tubulin patterns were the hardest for our current features and classifier, which suggested future work on designing better features to characterize their properties.

3.2 Feature Selection

To address the feature redundancy observed in our feature set, stepwise discriminant analysis (SDA) was conducted using the entire feature set containing 26 features. The max-win Gaussian-kernel SVM from above was used along with 10-fold cross validation to evaluate the effectiveness of the top-ranked features. The highest recall of 93% was achieved by using all 23 features returned from SDA. Although feature selection did not improve our original performance with the entire feature set, valuable insights can be obtained by inspecting the features ranked by SDA. Table 4 shows the top 15 features returned from SDA. The top 7 features were all Haralick texture features followed by the edge fraction feature (SLF7.9). Two more Haralick texture features were selected before the average object size feature (SLF1.3), two skeleton features (SLF7.81-82), and one edge feature (SLF7.11) were included. 9 out of the 10 top features returned from SDA were Haralick features and all 13 Haralick texture features were returned by SDA. The feature selection results along with the observed classification performance confirm the

Feature Set	Precision (%)	Recall (%)
13 Haralick texture features	88.8	89.0
13 Morphological features	76.6	75.8
Haralick + Morphological	93.1	93.2

Table 2. Evaluation of the two feature subsets and their combination using a DAG Gaussian-kernel SVM along with 10-fold cross validation.

True Class	Output of Classifier									
	DN	**ER**	**Gia**	**gpp**	**LA**	**Mit**	**Nuc**	**Act**	**TfR**	**Tub**
DNA	**100**	0	0	0	0	0	0	0	0	0
ER	0	**96**	0	0	0	0	0	0	4	0
Gia	0	0	**100**	0	0	0	0	0	0	0
Gpp	0	0	2	**98**	0	0	0	0	0	0
Lam	0	0	0	4	**94**	0	0	0	2	0
Mit	0	0	0	2	0	**96**	0	0	2	0
Nuc	0	0	0	0	0	0	**100**	0	0	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	4	0	0	2	4	4	0	**82**	4
Tub	0	4	0	0	2	4	0	0	8	**82**

Table 3. Confusion matrix for a max-win Gaussian-kernel SVM classifier using the entire feature set. The average precision and recall are both 94.8%.

strong discriminative ability of the Haralick feature subset for the purpose of analyzing subcellular images.

4. CONCLUSION

We described a successful approach to classifying subcellular patterns in multicell fluorescence microscope images. Better performance was achieved by classifying 2D multicell fluorescence microscope images than for single-cell images, which can be partially attributed to the greater amount of pattern information contained in multicell images. The feature subsets adapted from our previous single-cell image feature set worked successfully on multicell images. Haralick texture features demonstrated more discriminative capability than morphological features in both the classification and feature selection experiments. We are currently researching the effects of a variety of extensions to our approach, including analysis of multicell images that contain multiple subcellular patterns, and multicell images that contain overlapping cells and partial cells. We hope that our work on multicell images will extend smoothly to classifying subcellular patterns in tissue images without cell segmentation.

5. REFERENCES

[1] M. V. Boland, M. K. Markey, and R. F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," Cytometry, vol. 33, pp. 366-375, 1998.

SLF Index	Feature Description
SLF7.76	Haralick: difference entropy
SLF7.68	Haralick: correlation
SLF7.74	Haralick: entropy
SLF7.69	Haralick: sum of squares variation
SLF7.78	Haralick: info. measure of correlation 2
SLF7.70	Haralick: inverse difference moment
SLF7.72	Haralick: sum variance
SLF7.9	Morph: edge pixel fraction
SLF7.77	Haralick: info. measure of correlation 1
SLF7.71	Haralick: sum average
SLF1.3	Morph: average object size
SLF7.82	Morph: skeleton object fraction
SLF7.81	Morph: skeleton length to convex hull ratio
SLF7.11	Morph: max/min in the edge intensity histogram
SLF7.73	Haralick: sum entropy

Table 4. Top 15 features returned by stepwise discriminant analysis (SDA) after feature selection on the entire feature set.

[2] 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, vol. 17, pp. 1213-1223, 2001.

[3] K. Huang, M. Velliste, and R. F. Murphy, "Feature Reduction for Improved Recognition of Subcellular Location Patterns in Fluorescence Microscope Images," Proc SPIE, vol. 4962, pp. 307-318, 2003.

[4] M. Velliste and R. F. Murphy, "Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images," presented at 2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002), Bethesda, MD, USA, 2002.

[5] A. Garrido and N. de la Blanca, "Applying deformable templates for cell image segmentation," Pattern Recognition, vol. 33, pp. 821-832, May, 2000.

[6] S. Sclaroff and L. Liu, "Deformable shape detection and description via model-based region grouping," IEEE Trans. Pattern Anal. Machine Intell., vol. 23, pp. 475-489, May, 2001.

[7] Q. Yang and B. Parvin, "Harmonic cut and regularized centroid transform for localization of subcellular structures," IEEE Trans. Biomedical Engineering, vol. 50, pp. 469-475, April, 2003.

[8] R. F. Murphy, M. Velliste, and G. Porreca, "Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images," VLSI Sig Proc, vol. 35, pp. 311-321, 2003.

[9] V. Vapnik, Statistical Learning Theory. New York City: Wiley, 1998.

[10] U. Kressel, "Pairwise Classification and Support Vector Machines," in Advances in Kernel Methods - Support Vector Learning, vol. 15, C. B. a. A. J. S. B. Scholkopf, Ed. Cambridge, Massachusetts: MIT Press, 1999.

[11] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification," Adv Neural Inform Proc Systems, vol. 12, pp. 547-553, 2000.