33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Features for Cells and Nuclei Classification

Song Liu, Piyushkumar A. Mundra, and Jagath C. Rajapakse

***Abstract*—The performance of automated analysis of cellular images is heavily influenced by the features that characterize cells or cell nuclei. In this paper, an exhaustive set of features including morphological, topological, and texture features are explored to determine the optimal features for classification of cells and cell nuclei. The optimal subset of features are obtained using popular feature selection methods. The results of feature selection indicate that Zernike moment, Daubechies wavelets, and Gabor wavelets give the most important features for the classification of cells or cell nuclei in fluorescent microscopy images.**

I. INTRODUCTION

Microscopic visualization of cell morphologies and their behaviors are often used as aides in the diagnosis of many diseases. Identification of abnormal cells is usually performed by staining the cells with various compounds and then imaging using fluorescent microscopy to determine cell states. For example, microscopic images taken from cancerous tissues or cells undergoing cell-cycle changes have been used in the diagnosis of cancer [1] [2]. Using such images, pathologists make judgements on tissue status based on their experiences. Such manual decisions are subjected to observer variability and affect the reliability of the diagnosis.

To circumvent drawbacks of manual classification of cells, it is important to develop computational tools that give quantitative measures for classification of cell types. Such informatics tools first separate individual cells or cytological components from the extracellular space. Subsequently, different categories of features including Haralick [3] and wavelets [4] are explored to characterize the individual cells. When features characterizing different properties of cells are combined properly, the accuracy of classification cells generally improves. Though a variety of features can be estimated from cells, it is vital to determine the most relevant features that give the highest classification accuracy because all features are not equally relevant and redundancy among the features usually affects the performance. The selection of features can be achieved using filter, wrapper, and embedded methods [5]. Feature selection methods aim to detect a set of features, which leads to optimal classification of cells or tissue classes.

In this work, an exhaustive set of image features of cells/nuclei are explored to find the optimal features for automated recognition of cell categories. The focus is mining of various features like morphological features, Zernike moments, Haralick texture features, wavelet features, run length features, etc., which highlight different properties of cells. Morphological features characterize the size of the objects in the cells, the intensity of edges, or the contour of the cells [6] [2]. Zernike moment features are computed from a set of Zernike polynomials and are good shape descriptors of cells [6]. Haralick texture features define 13 features describing the grey level relationship of the pixels of an image and capture the texture information of an image [3] [6]. The Daubechies wavelet feature set [7] and Gabor wavelet features [8] both give interpretation of cells in a multi-resolution way. Similar to Haralick features the gray level run length method is based on computing the number of gray level runs of various lengths [9] and compute textures from the cells.

To select the optimal sets of features for better recognition of cells, we explore four widely used feature selection methods: F-score [10], Kruskal-Wallis (KW)-test [11], minimum redundancy maximum relevancy (MRMR) criterion [12], and support vector machine - recursive feature elemination (SVM-RFE) method [13] [14]. Our experiments indicate that Zernike moments, Daubechies wavelet features, and Gabor wavelet features are three most important features for cell/nuclei classification. SVM-RFE is the best feature selection method that generated the smallest feature subset while maintaining a good classification performance.

II. FEATURES

Consider a 2-D tissue image $f : \Omega \to \mathbb{R}$ where $f(z)$ denotes the image intensity at pixel $z \in \Omega \subset \mathbb{N}^2$ and Ω is the image domain. Suppose J number of features from I cells were extracted from the image, so the dataset $D = \{x_{ij}\}_{i=1,j=1}^{I,J}$ of features would be available where x_{ij} refers to the value of feature j of cell i. The vector $x_i = (x_{ij})_{j=1}^{J}$ denotes the feature vector representing cell i and $x_j = (x_{ij})_{i=1}^{I}$ denotes the features j across all the cells. Let the label or class of cell i be $y_i \in l \in \Gamma = \{l\}_{l=1}^{L}$ where L is the number of cell classes.

A. Morphological Features

Morphological features describe various characteristics of fluorescent objects and edges in the cellular image as well as the entire cells. These features usually characterize the size of the fluorescent objects, the intensity of edges, or the contour of the cells [6]. Sixteen morphological features related to fluorescent objects or edges in the fluorescent images were obtained. These features describe, for example, the number of fluorescent objects or the edge direction homogeneity.

B. Zernike Features

Zernike moments are calculated using an orthogonal basis of Zernike polynomials [6]. Zernike moments are the projections of an image onto the orthogonal basis functions. They are the magnitude of a set of orthogonal complex moments that are spatially and rotationally invariant. We calculated Zernike

Song Liu, Piyushkumar A. Mundra, and Jagath C. Rajapakse are with BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological university, Singapore (email: {y060101, MUND0001}@e.ntu.edu.sg, {asjagath}@ntu.edu.sg).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

moment features by using the Zernike polynomials with orders ranging from 2-20. The results indicate that there is no significant difference of the classification using just Zernike moments with order over 12. In this case, we calculated 49 Zernike features with order 12.

C. Haralick Features

Haralick features are texture features computed from the grey level co-occurrence matrix of the image [3]. Suppose a given image f has N grey levels and its grey level co-occurrence matrix is given by $\{q(i,j)\}_{i=1,j=1}^{N,N}$ matrix in which co-occurence $q(i,j)$ gives the frequency with which two neighboring pixels having grey level i and grey level j. As a result, 13 different texture features were obtained after averaging.

D. Daubechies Wavelet Features

Wavelet packets are a generalization of orthonormal and compactly supported wavelets [7]. The coefficients of decomposition serve as distinct features of the original image. Initially, the image was decomposed up to 4th level and 30 wavelet features were extracted to represent the frequency information that best represents the image. Further, the image was decomposed up to level 10 and the average energies of three high-frequency images at each level were used as features. In the latter case, a total set of 300 features were generated.

E. Gabor Wavelet Features

Information captured by nonorthogonal Gabor wavelets is mostly the derivative information of an image such as edges [8]. Gabor wavelets are a set of basis functions generated through dilation and rotation of a mother Gabor wavelet. The input image was convolved with a Gabor filter at a specific scale and at a particular orientation. The mean and standard deviation of the responses are taken as texture features. We have used five different scales and six different orientations, yielding a total of 60 Gabor texture features.

F. Run Length Features

Run length features were extracted from the grey level run of the image. A grey level run is a connected set of pixels in a specific direction having the same grey values [9]. Grey level runs can be used to characterize the spatial variation of pixel values in an image. Once the run-length matrices are calculated along each direction, several texture descriptors are calculated to capture texture properties and differentiate different textures. Finally, 22 run length features were obtained.

III. Feature Selection

A. F score

F-score is used for ranking features in multi-class problems and is given by the ratio of between-class sum of squares to within class sum of squares [10]. It presumes independence among features and Gaussianity of class distributions. Hence, it is computationally efficient but suffers from redundancy and non-Gaussianity of data.

B. KW-test

Kruskal-Wallis (KW) test is nonparametric feature selection method [11], which is based on the distributions of ranks in different classes. It does not assume Gaussianity of class distributions and is robust to variations of the data.

C. MRMR

Filter criteria such as F-test and KW-test are solely based on their relevance with respect to the class labels and suffer from redundancy among the features. Minimum redundancy maximum relevancy (MRMR) criterion is therefore proposed to select features that are maximally relevant to the prediction of classes while keeping the redundancy among the features to a minimum level [12]. In this paper, we have used ratio of F-score (relevancy) to Pearsons correlation coefficient (redundancy among features) to compute MRMR criterion.

D. SVM RFE

SVM weights represent the importance of the corresponding input or feature in its classification. Using all the features to begin, SVM-RFE eliminates features recursively in a backward elimination manner for identification of relevant features [13]. The standard SVM-RFE algorithm was originally proposed for two-class classification problem and are extended for multiclasses [14] .

IV. Experiments and Results

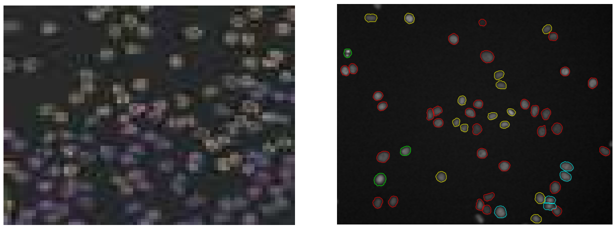
Cells in all the image datasets were first segmented into individual cells by using the multi-phase level sets [15], followed by a marker-controlled watershed algorithm [16]. After that, various types of features described above were extracted over every individual cells to represent cells. Using different feature selection methods, the optimal feature set was selected in order to maximize the classification accuracy.

A. Synthesized Cellular Images

Synthetic cellular images were generated to simulate 2D fluorescent images of cells. The simulated cells consist of two main compartments: cytoplasm and nuclei. Morphologies and textures of the cytoplasm and nuclei were generated using the models described in [17] (Each image contain cells from the same class). In addition, multiplicative Poison noise and additive Gaussian noise were added to represent the variations in photon emission and the scanner, respectively. Five different classes of synthesized cells were generated for the experiments. There are totally 4773 cells for the classification.

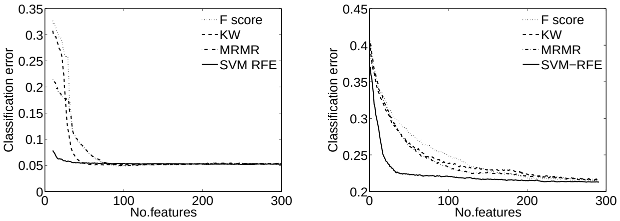
B. Cell Cycle Images

Publicly available 2D image sequence generated for DCellIQ project were used to test the efficiency of the proposed method (http://www.cbi-tmhs.org/Dcelliq/index.html). The sequences consist of 100 images of Hela cells obtained from time-lapsed microscopy. The aim was to identify cells belonging to 4 different phases of cell-cycle. The ground truth



(a) (b)

Fig. 1. Illustration of image datasets: (a) synthesized cell image, cells in the image are from the same class, (b) cell cycle image, each image contain multi-class of cells, different colored contours represent different classes of cells.



(a) (b)

Fig. 2. The influence of number of features over classification accuracy: (a) synthesized cell data, (b) cell cycle data.

of cell-cycle data was provided along with the data. There are totally 9103 cells for the classification.

C. Performance Evaluation

To reduce computational time, the feature ranking and performance of the classification was evaluated using 100 times bootstrapping. During each run, 100 samples were randomly selected for each class as training data and features were ranked using various techniques. The remaining samples from original datasets were then used as corresponding test datasets. The methods were evaluated on test samples with the selected features. Features were normalized to zero mean and unit variance before ranking and testing.

The final selection of features was done by ranking the features according their emerging frequencies, i.e., how frequently a particular feature appears in the top 100 features during bootstrapping. The final classification errors are calculated as an average over the classification during 100 times bootstrapping. LIBSVM - version 2.84 software [18] was used for implementation of feature selection methods and classification with SVM.

D. Results

Due to 100 bootstrapped testing, 100 feature rank lists for each method were obtained. Aiming to analyze the contribution of each class of features, the important features with selection frequency larger than 50 were listed in Table I-II. Here, the classification is calculated using features with selection frequency larger than threshold of 50, which is close to the turn point of the classification error curve for SVM-RFE in Figure 2. As seen, the number of selected texture features exceeds the number of selected morphological features. However, comparing the ratio between the number of selected features to the total number, it is found that morphological features have higher such ratio than the texture features. This reflects the fact that the selected morphological features play important roles and can be attributed to the critical inter-class morphology difference among cells in cell cycle images. Among the features, Zernike moments, Daubechies wavelets, and Gabor wavelet features were three most important features. The remaining features had different contributions to classification results. Some types of features reveal poor contribution to the classification as their emerging frequencies were less than 50.

Zernike moments, Daubechies wavelets, and Gabor wavelets represent an image with a set of orthogonal basis functions or polynomials but extracting different properties: shape, spatial scale, and spatial frequency, respectively. The cell-cycle data consists of nuclei while the synthesized data consists of whole cells representing both nuclei and cytoplasm. As a result, two datasets yield different sets of features. For both datasets, texture seems to play an important role in the recognition of cells.

As shown in Figure 2, SVM-RFE outperformed other feature selection methods by selecting the smallest subset of features with the highest classification accuracy for the identification of cells. SVM-RFE is a wrapper approach that incorporates features importance, using SVM weights. On the other hand, filter approaches such as F-score, KW-test, and MRMR, neglect the significances of features with respect to the classifier. As a result, these methods need a larger feature subset compared to SVM-RFE. Comparing filter methods used in this study, MRMR penalizes for redundant features. With respect to redundancy, SVM-RFE penalizes for redundant features and hence less number of wavelet packet features were selected compared to filter approaches. unlike SVM-RFE, MRMR is a filter method which does not include the classifier characteristics, in feature selection. It generally results in poor classification performance compared to SVM-RFE.

TABLE I
SELECTED FEATURES FOR CLASSIFICATION OF CELLS OF CELL CYCLE DATA

Cycle	Morph	Zern	Wavser	Gabor	Accu(%)
F score	2	2	94	1	93.71
KW	2	8	86	5	94.84
MRMR	4	10	71	8	93.06
SVM-RFE	7	20	27	21	94.58

TABLE II
SELECTED FEATURES FOR CLASSIFICATION OF CELLS OF SYNTHETIC DATA

Synthesized	Morph	Zern	Hara	Wavser	Gabor	Run	Accu(%)
F score	0	6	11	29	31	22	71.98
KW	0	6	5	48	41	0	73.43
MRMR	0	6	11	30	31	22	73.69
SVM-RFE	5	7	0	23	30	0	77.63

V. CONCLUSIONS

Our experiments revealed that Zernike moments, Daubechies wavelets features, and Gabor wavelets features are three most important features for cell/nuclei classification. Though these methods characterize the images with a set of orthogonal basis functions, they extract different kinds of features of cells: Zernike moments for shape features, Daubechies wavelets for spatial scales, and Gabor wavelets for spatial frequencies of an image of a cell. SVM-RFE was the best feature selection method that generated the smallest feature subset while rendering the highest performance. Feature selection of cell classification is a vital part of designing dedicated and efficient informatics solutions for cell-based disease diagnosis and drug design.

REFERENCES

[1] A. Basavanhally, S. Ganesan, S. Agner, J. Monaco, and et al., “Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology,” *IEEE Transactions Biomedical Engineering*, vol. 57, no. 3, pp. 642–653, 2010.

[2] M. Wang, X. Zhou, F. Li, and J. Huckins, “Novel cell segmentation and online svm for cell cycle phase identification in automated microscopy,” *Bioinformatics*, vol. 24, no. 1, pp. 94–101, 2008.

[3] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, and et al., “Automated grading of prostate cancer using architectural and textural image features,” in *In Biomedical Imaging: From Nano to Macro. ISBI 2008. 5th IEEE International Symposium*, 2007.

[4] R. Murphy, “Location proteomics: a systems approach to subcellular location,” *Biochemical Society Transactions*, vol. 33, pp. 535–538, 2005.

[5] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[6] M. 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, no. 12, pp. 1213–1223, 2001.

[7] K. Huang and R. Murphy, “From quantitative microscopy to automated image understanding,” *Journal of Biomedical Optics*, vol. 9, no. 5, pp. 893–912, 2004.

[8] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and et al., “Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features,” in *In Biomedical Imaging: From Nano to Macro. ISBI 2008. 5th IEEE International Symposium*, 2008.

[9] X. Tang, “Texture information in run-length matrices,” *IEEE Transactions Image Processing*, vol. 7, no. 12, pp. 1602–1609, 1998.

[10] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *J American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.

[11] D. Chen, Z. Liu, X. Ma, and D. Hua, “Selecting genes by test statistics,” *J. Biomedicine and Biotechnology*, vol. 2, pp. 132–138, 2005.

[12] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *J Bioinformatics Computational Biology*, vol. 3, pp. 185–205, 2005.

[13] I. Guyon, J. Weston, S. Barhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, pp. 389–422, 2002.

[14] X. Zhou and T. P. Tuck, “MSVM-RFE: Extensions of svm-rfe for multiclass gene selection on dna microarray data,” *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[15] L. Vese and T. Chan, “A multiphase level set framework for image segmentation using the mumford and shah model,” *International Journal of Computer Vision*, vol. 50, no. 3, pp. 271–293, 2002.

[16] J. Chen and J. Rajapakse, “Segmentation of clustered nuclei with shape markers and marking function,” *IEEE Transactions Biomedical Engineering*, vol. 56, no. 3, pp. 741–748, 2009.

[17] A. Lehmussola, P. Ruusuvuori, J. Selinummi, and et al., “Computational framework for simulating fluorescence microscope images with cell populations,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 7, pp. 1010–1016, 2007.

[18] C. Chang and C. Lin, “Libsvm: A library for support vector machines,” *www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.