Breast cancer mitosis detection in histopathological images with spatial feature extraction

Abdülkadir Albayrak, Gökhan Bilgin

Abdülkadir Albayrak, Gökhan Bilgin, "Breast cancer mitosis detection in histopathological images with spatial feature extraction," Proc. SPIE 9067, Sixth International Conference on Machine Vision (ICMV 2013), 90670L (24 December 2013); doi: 10.1117/12.2050050

Event: Sixth International Conference on Machine Vision (ICMV 13), 2013, London, United Kingdom

Breast Cancer Mitosis Detection in Histopathological Images with Spatial Feature Extraction

Abdülkadir Albayrak and Gökhan Bilgin
Department of Computer Engineering, Yildiz Technical University, 34220 Istanbul, Turkey
{albayrak, gbilgin}@yildiz.edu.tr

ABSTRACT

In this work, cellular mitosis detection in histopathological images has been investigated. Mitosis detection is very expensive and time consuming process. Development of digital imaging in pathology has enabled reasonable and effective solution to this problem. Segmentation of digital images provides easier analysis of cell structures in histopathological data. To differentiate normal and mitotic cells in histopathological images, feature extraction step is very crucial step for the system accuracy. A mitotic cell has more distinctive textural dissimilarities than the other normal cells. Hence, it is important to incorporate spatial information in feature extraction or in post-processing steps. As a main part of this study, Haralick texture descriptor has been proposed with different spatial window sizes in RGB and La*b* color spaces. So, spatial dependencies of normal and mitotic cellular pixels can be evaluated within different pixel neighborhoods. Extracted features are compared with various sample sizes by Support Vector Machines using k-fold cross validation method. According to the represented results, it has been shown that separation accuracy on mitotic and non-mitotic cellular pixels gets better with the increasing size of spatial window.

Keywords: Histopathological images, mitosis detection, Haralick features, spatial information extraction.

1. INTRODUCTION

During the last decade, digital analysis of histological images has become very important for early diagnosis of every types of cancer. Histologists mainly focus on structural and morphological irregularities of cellular structures in the evaluation of histological images. According to these irregularities several grading systems have been developed for determination the level of malignance of cancerous tumors. Various grading systems are used in the diagnosis and prognosis stages of histopathology. Gleason, Scarff-Bloom-Richardson, Elston-Ellis grading systems are very popular grading systems for regular diagnosis [1-3]. Cancerous mitotic activity is one of the most determinative signs for invasive breast cancer development. It is defined as number of mitotic event in a specified region of interest in tissue. Mitotic events are assessed within all grading systems as mentioned before.

In an automated digital histopathological image analysis, mitosis detection is assumed as a task of 'mitotic cellular pattern recognition'. Hence, similar to a pattern recognition system, feature extraction step is very crucial step for system accuracy. Extracted features from histopathological images are highly related to the cell morphology. For this reason, extraction of individual cell via segmentation becomes very important in automated image analysis. In general, a thresholding method is used for elimination of background effects before segmentation. Then, the cellular structures are separated from background with a clustering algorithm including thresholding step or all alone. Afterwards, distinguishable features are tried to be extracted for every individual cell structure. Finally, a classification approach is applied which realizes maximum discrimination of related features.

Many techniques on cellular segmentation in histopathological images have been proposed in the literature. An automated cell segmentation of nucleuses is proposed using the gradient magnitude and pixel direction of nucleuses in [4]. Morphological segmentation of histology images has been studied in [5]. Segmentation of microscopic cell images using adaptive eigenfilters was introduced in [6]. Furthermore, there may be some problematic issues while applying segmentation based operations. Mathematical morphological operations can be used to overcome these problems as in [7]. In a proposed study, morphological opening, closing, erosion, and dilation operations are implemented to remove noise after segmentation as a post-processing step in [8]. Unsupervised algorithms, as mentioned above, use statistical information while segmenting an image. In some cases segmentation cannot be done successfully because of the irregularity of color distribution in histopathological images [9]. In some cases supervised segmentation may be helpful

Sixth International Conference on Machine Vision (ICMV 2013), edited by Antanas Verikas,
Branislav Vuksanovic, Jianhong Zhou, Proc. of SPIE Vol. 9067, 90670L · © 2013
SPIE · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2050050

to segment the images with easily clearable noise or erasable unnecessary structures before feature extraction process. Getting *a priori* knowledge from ground truth of histopathological images makes segmentation easy for supervised segmentation algorithms [10-11].

In histopathological image segmentation, incorporating spatial information is more valuable than pure pixel information for enhanced analysis. Exploitation of the textural information, as an important spatial information source, helps us to reveal distinguishable features between cellular and other background structures in histopathological images. Classification of the follicular lymphoma using a novel color texture analysis method is proposed to classify the image into low or high grade in [12]. A Gabor filter based feature extraction algorithm with adding textural information of the histopathological images is introduced in [13]. A graph-based multi-resolution method for mitosis detection in Breast Cancer Histological Images has been proposed in [14]. Briefly, in a given resolution, the image was simplified by discrete regularization and then two-means clustering is performed to specific region that were segmented at the previous resolution. The clustered region was refined at each resolution step and possible mitotic regions are defined at the last step. In another work, it is proposed to find a strong mitosis detection approach by using Scale Invariant Feature Transform (SIFT) [15]. Formal variability of the interested regions are studied for a classification system of tumor tissues of breast cancer images in [16].

In this study, Haralick texture features descriptor method [17] is exploited and evaluated for robust mitosis detection in breast cancer histopathological images. In the first step, gray level co-occurrence matrices (GLCM) are computed for each bands (or layers) in RGB and La*b* color-space. Then GLCM are used in Haralick algorithm to extract the features of the pixel of interest with different window sizes. With the changing windows size, spatial dependencies and relationships between pixels can be revealed to extract meaningful features for a given pixel. At the last step, SVM as a supervised classification algorithm has been applied to classify the extracted features. The effects of different window sizes and the color distributions in RGB and La*b*color spaces are also compared in this work.

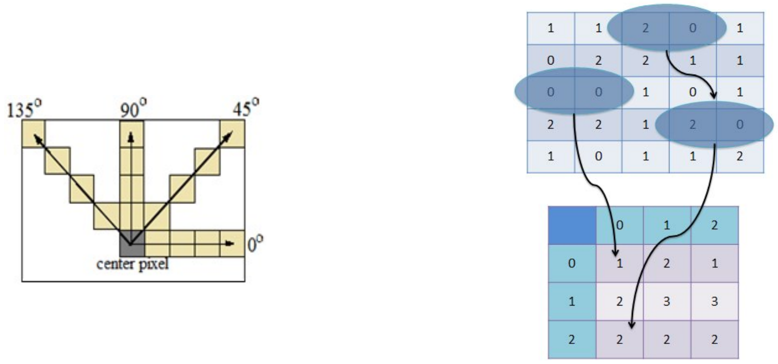
The rest of the paper is organized as follows. Section II provides the methodology that is described briefly. In Section III, the experimental results are presented and discussed. Finally, in Section IV we conclude this paper with final remarks and comments.

2. SPATIAL TEXTURE DESCRIPTOR

Texture analysis is considered as one of the most important spatial characteristics in image processing techniques. The similarity of the objects or regions can be found by using textural information of the specified areas. In this study Haralick feature descriptor is used for feature extraction with exploiting spatial dependencies and textural relationships. The co-occurrence probabilities are computed with gray level co-occurrence matrix for input of Haralick method.

2.1 Gray Level Co-occurrence Matrix (GLCM)

A gray-level co-occurrence matrix is described by calculating the adjacency of a pixel i to another pixel j in a given image. In this study, four different matrices are obtained from different orientations which are horizontally (P^0), vertically (P^{90}), center pixel to top-right (P^{45}), and center to top-left (P^{135}). Fig.1 (a) represents directional analysis of a center (referenced) pixel which is randomly selected from image.



1	1	2	0	1
0	2	2	1	1
0	0	1	0	1
2	2	1	2	0
1	0	1	1	2

	0	1	2
0	1	2	1
1	2	3	3
2	2	2	2

(a)

(b)

Figure 1. a.) Orientations of center (referenced) pixel while GLCM is obtained, b.) Example for obtaining a Gray Level Co-occurrence Matrix

Fig.1 (b) represents a 5x5 given image patch and a 3x3 GLCM of the given image. Since the values in 5x5 input image are 0, 1, and 2, a three level scale will be used. In 5x5 image there is only one state that 1 is followed by 1 and two states that 2 is followed by 0 in the given image part. In parallel, when the offset is enlarged with different coordinates the co-occurrence matrix is updated according to the offset values specified. Fig. 1(b) represents a co-occurrence matrix with one distance, eight gray-level numbers of neighbors, and the 0^o orientation.

In the feature extraction step some statistical features which are calculated from GLCM are extracted. These statistical features are consist of energy, contrast, homogeneity, entrophy, correlation and variance. These statistical features can be summarized as follows: $P[i,j]$ is the GLCM 3x3 window size represented in Fig.1 (b) because the 5x5 given images values 0, 1, and 2.

$$energy = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{P^2[i,j]\} \tag{1}$$

where $P[i,j]$ is the gray-level co-occurrence matrix, N is the gray level related to the image type whether it is binary or intensity. Index i denotes the center pixel while index j is used for neighbor pixels.

$$contrast = \sum_{i=1}^{N}\sum_{j=1}^{N}\{(i-j)^2 P[i,j] \tag{2}$$

$$homogeneity = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\frac{P[i,j]}{1+|i-j|} \tag{3}$$

$$entrophy = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P[i,j]xlog(P[i,j]) \tag{4}$$

$$correlation = \sum_{j=0}^{N-1}\sum_{j=0}^{N-1}\frac{\{i\ x\ j\}x\ P[i,j] - \{\mu_x\ x\ \mu_y\}}{\sigma_i x\ \sigma_i} \tag{5}$$

where μ and σ are the mean and standard deviation values of P respectively,

$$variance = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(i-\mu)^2 P[i,j] \tag{6}$$

3. EXPERIMENTAL SETUP

In this study, ICPR2012 conference contest dataset, “Mitosis Detection in Breast Cancer Histopathological Images”, is used for the evaluation. The dataset is provided by two experienced pathologists for the contest. The dataset consists of 5 breast cancer biopsy slides which are stained with hematoxylin and eosin (H&E). Ten high power fields (HPF) with area of 512x512 μm^2 are selected by the pathologists.

Totally, 50 HPFs are scanned by three different scanners: A, H, and a multispectral M scanner. Scanner A has a resolution of 0.2456 μm^2 per pixel. Scanner H has resolution of 0.2273 μm^2 horizontal and 0.22753 μm^2 vertical resolutions per pixel. Multispectral M has the best resolution of 0.185 μm^2 per pixel. There are 226 mitosis cases in total. In this study, only H scanner images are used as the training and testing data to find mitosis cases.

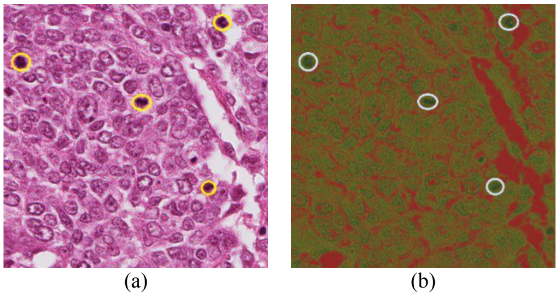


Figure 2. (a) A sample image segment selected from database (b) La*b* channel representation of same sample image.

The study is implemented all in Matlab^{TM} 2012a platform. Both the RGB and La*b* color spaces are used for pixel-wise classification. The comparison of RGB and La*b* color spaces are represented separately in the following parts of this paper.

In the first step, the co-occurrence matrices of randomly selected pixels from mitotic and non-mitotic pixel areas are obtained in the original RGB and La*b* color spaces. Created matrices created with 4 different window sizes which define 3x3, 5x5, 7x7, and 9x9 surrounding areas of center referenced pixel. The co-occurrence matrices are then normalized to 8x8 matrices which include the pixel relationship within the neighbor pixels. For example, if the offset refers only to the next pixel of the referenced pixel (in 3x3 windows size), then the co-occurrence matrix includes number of neighbor pixel comes after the referenced pixel. In that study, the offset has been chosen on the vertical, horizontal and diagonal neighbor pixels within 0^{o}, 45^{o}, 90^{o}, and 135^{o} degrees. Fig. 3 represents computation of the co-occurrence matrices for different randomly selected pixels.

After getting the co-occurrence matrices, fast calculation of Haralick feature descriptor algorithm was applied to the selected pixels as in [18]. In Haralick implementation, 13 different features are extracted. Those features are energy, correlation, sum of variances, inverse different moment (homogeneity), sum average, entropy, sum of entropy, sum of variance, difference of variance, contrast, difference of entropy and two information measures of correlation. For a pixel that has 9x9 window size, total 624 features are extracted.

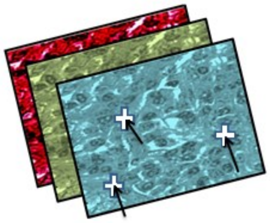


Figure 3. Obtaining Gray Level Co-occurrence Matrix of selected pixels.

Radial Basis Function (RBF) kernel based SVM algorithm is used for the classification process. To solve optimization problem in the training phase of SVM, sequential minimal optimization (SMO) algorithm was considered [19]. The best results are obtained with tolerance parameter value of 0,001, cost function value of 10 and gamma value of 0,04 gamma for RBF kernel with 5-fold cross-validation. Table 1 represents the classification accuracies of several sub-datasets created with different number of mitotic pixel samples (500, 1000, 2500 and 5000) chosen from all histopathological slides. Those subsets consist of specified number of randomly selected original mitosis pixels (in column 1) and several times ($\mathbf{xM}$; $\mathrm{M} = 1,2,3,5,10$ and 20) larger number of non-mitosis pixels extracted from original images using groundtruth information. M can be called as mixture multiplier. For example, if the number of mitotic pixel samples is 2500 and x3, there are 2500 mitotic and 3x2500 non-mitotic pixels in that sub-dataset (total 10000 pixels).

Table 1 represents the accuracies of the same pixels in RGB and La*b* color space. The classification accuracies of the original pixels both in RGB and La*b* spaces are almost the same (La*b* is a little bit higher).

In Table 2, the classification accuracies are represented for Haralick implementation of same pixels shown in Table 1 with different window sizes. In all sub-datasets, the accuracy increases when the number of mitotic samples and also mixture multiplier (xM) increase. Accuracies of Haralick implemented pixels with 3x3, 5x5, 7x7 and 9x9 window sizes get better as the window size and mixture multipliers increases compared to Table 1. Table 3 shows the results of 3x3, 5x5, 7x7 and 9x9 window sizes respectively in La*b* color space. The accuracies of Haralick implemented sub-datasets with different window sizes in RGB and La*b* color spaces are almost the same with very small changes. It can be said that color space conversion has very small effects on classification accuracies.

Table 1. Accuracies in percentages for RGB and LAB in original feature space

Original	# of samples	x1	x2	x3	x5	x10	x20
RGB	500	89.00	90.60	90.35	92.80	95.22	96.78
	1000	88.95	90.46	91.15	93.15	95.31	96.87
	2500	89.82	90.37	91.29	92.99	95.27	97.05
	5000	89.34	89.82	90.92	92.98	95.14	96.99
La*b*	500	89.60	90.40	90.40	92.93	95.18	96.77
	1000	89.20	90.73	91.03	93.32	95.36	96.89
	2500	89.68	90.53	91.38	93.05	95.31	97.07
	5000	89.39	89.77	90.98	93.00	95.16	97.02

Table 2. Accuracies in percentages for RGB with different window sizes

w.s*	# of samples	x1	x2	x3	x5	x10	x20	w.s*	# of samples	x1	x2	x3	x5	x10	x20
3x3	500	93.10	91.60	93.10	94.70	96.30	97.76	7x7	500	94.5	95.6	95.55	97	98.25	98.76
	1000	91.35	92.70	93.15	94.91	96.65	97.84		1000	95.85	95.73	97.11	97.41	98.30	98.85
	2500	93.44	93.01	93.98	96.11	96.66	97.87		2500	96.68	96.47	97.01	97.62	98.35	99.07
	5000	92.69	93.48	94.08	94.99	96.67	97.89		5000	96.96	97.10	97.64	98.10	98.38	99.17
5x5	500	93.3	93.26	94.90	95.83	97.4	98.45	9x9	500	94.6	95.73	95.8	96.6	98.14	98.91
	1000	93.96	93.97	95.33	95.97	97.22	98.41		1000	95.8	95.93	97.12	97.16	98.30	98.88
	2500	95.12	94.96	95.46	96.25	97.52	99.05		2500	96.32	96.51	97.06	97.72	98.44	99.09
	5000	95.17	95.61	96.08	96.55	97.65	98.58		5000	97.06	98.41	97.58	98.04	98.56	99.15

w.s*: window size

Table 3. Accuracies in percentages for La*b* with different window sizes

w.s*	# of samples	x1	x2	x3	x5	x10	x20	w.s*	# of samples	x1	x2	x3	x5	x10	x20
3x3	500	93.30	90.93	93.00	94.63	96.38	97.77	7x7	500	94.90	95.60	96.30	96.93	98.22	98.90
	1000	92.40	92.57	93.23	94.73	96.58	97.82		1000	92.40	92.57	93.23	94.73	97.20	98.94
	2500	93.40	92.95	94.06	94.87	96.66	97.90		2500	93.40	92.95	94.06	94.87	96.66	98.97
	5000	92.88	92.90	94.38	94.98	96.98	97.89		5000	92.88	93.01	94.11	94.98	96.70	99.01
5x5	500	93.60	93.33	94.55	95.70	97.42	98.37	9x9	500	95.50	97.07	97.15	97.47	98.71	99.22
	1000	93.35	94.40	95.10	96.03	97.30	98.39		1000	97.15	97.17	97.70	98.18	98.93	99.28
	2500	94.98	94.99	95.50	96.33	97.54	98.52		2500	97.84	97.80	98.51	98.57	98.75	99.32
	5000	95.08	95.03	95.63	96.66	97.60	98.58		5000	98.25	97.94	98.87	99.03	99.07	99.30

w.s*: window size

4. CONCLUSION

In this study, mitosis detection in histopathological images using the spatial information has been proposed. Haralick implementation of randomly selected mitotic and non-mitotic pixels with different window sizes has been compared. The SVM classification accuracies with respect to total number of samples of mitotic and non-mitotic pixels (which were mixed with different ratios to mitotic pixels) have been compared. The accuracy results are shown in tables separately in each RGB and La*b* color spaces. According to the represented results, it has been shown that accuracy of separation of mitotic and non-mitotic cellular pixels gets better with the increasing size of spatial window.

REFERENCES

[1] G. Eason, B. Noble and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,"Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, (1955)

[2] H. J. G. Bloom and W. W. Richardson, "Histological grading and prognosis in breast cancer,"Br. J Cancer, vol. 11, pp. 359-377, (1957).

[3] C.W. Elston and I.O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up," Histopathology, vol. 19, pp 403-410, (1991).

[4] K. Nandy, P.R. Gudla and S.J. Lockett,"Automatic segmentation of cell nuclei in 2D using dynamic programming,"Proceedings of Second Workshop on Microsopic Image Analysis with Applications in Biology, Piscataway, NJ, USA, (2007)

[5] A. Nedzved, S. Ablameyko and I. Pitas,"Morphological segmentation of histology cell images," in: Proceedings of the International Conference on Pattern Recognition, vol. 1, pp. 500{503, Washington DC, USA, (2000)

[6] S. Kumar, S. Ong, S. Ranganath, F. Che and T. Ong ,"Segmentation of microscope cell images via adaptive eigenfilters," in Int'l. Conference on Image Processing, pp. 135-138, (2004)

[7] V. K. Awasthi , W. Doolitle , G. Parulkar and J.G. Mc Nally,"Cell Tracking using a Distributed Algorithm for 3D Image Segmentation,"Bio Imaging, (1994)

[8] D. Anoraganingrum,"Cell segmentation with median filter and mathematical morphology operation,"In: Proceedings of 1999 International Conference on Image Analysis and Processing, pp. 1043–1046, (1999)

[9] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," Comput. Vis. Image Underst., vol. 110, no. 2, pp.260 -280, 2008.

[10] H. Kong, M. Gurcan, K. Boussaid, "Partitioning histopathological Images: an integrated framework for supervised color-texture segmentation and cell splitting, "IEEE Transactions on Medical Imaging, issue 99, pp.1-18, (2011).

[11] B. Meftah, O. Lezoray, M. Lecluse and A. Benyettou,"Cell microscopic segmentation with spiking neuron networks,"Proc. 20th Int Conf. on Artificial Neural Networks, pp. 117–126, (2010)

[12] O. Sertel, J. Kong, G. Lozanski, A. Shanaah, U. Catalyurek and J. Saltz,"Texture classification using non-linear color quantization: Application to histopathological image analysis," In Proc. of the IEEE int. conf. on acoustics, speech, and signal processing (ICASSP), pp. 597–600, Las Vegas, (2008)

[13] S. Doyle , S. Agner , A. Madabhushi, M. Feldman and J. Tomaszewski,"Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features,"Proc. 5th IEEE Int. Symp. Biomed. Imag.: From Nano to Macro, pp.496 -499, (2008)

[14] V. Roullier, O. Lezoray, V-T. Ta and A. Elmoataz,"Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization,"Computer medical imaging and graph, (2011).

[15] H. Irshad, S. Jalali, L. Roux, D. Racoceanu, L.J. Hwee, G. L. Naour and F. Capron,"Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach," Journal of Pathology Informatics,"Volume 4, Issue 2, p. 12, (2013).

[16] L. E. George and K. H. Sager,"Breast cancer diagnosis using multi-fractal dimension spectra,"Proc. IEEE Int. Conf. Signal Process. Communication, pp.592 -595, (2007)

[17] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification,"IEEE Transactions on Systems, Man and Cybernetics, (1979)

[18] E. Miyamoto and T. Merryman, "Fast calculation of Haralick texture features," Human Computer Interaction Institute, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, 2011.

[19] J. Platt., "Sequetial minimal optimization: A fast algorithm for training support vector machines." In Technical Report MST-TR-98-14. Microsoft Research, 1998.