A Computer-aided System for Discriminating Normal from Cancerous Regions in IHC Liver Cancer Tissue Images Using K-means Clustering*

R. M. CHEN[1], Y. J. WU, S. R. JHUANG, M. H. HSIEH, C. L. KUO, Y. L. MA
Department of Computer Science and Information Engineering
National University of Tainan
33 Sec. 2, Shulin St., Tainan 70005
TAIWAN
[1]rmchen@mail.nutn.edu.tw

R. M. HU[2] and JEFFREY J. P. TSAI[3]
Department of Biomedical Informatics
Asia University
500 Lioufeng Rd., Wufeng, Taichung 41354
TAIWAN
[2]rmhu@asia.edu.tw [3]jjptsai@gmail.com

Abstract: - Immunohistochemistry (IHC) is a well established imaging technique that can be exploited to detect whether the target antigen exists in tissue sections or not in order to discriminate between the cancerous and normal regions in a cancer tissue specimen. The intensity of immuno-stained protein in normal and cancerous regions can be compared to detect the gene status in sample tissues. In this paper, we address the problem of identifying the differential expression of marker protein on cancerous and normal regions in an IHC liver cancer tissue image. We present an improved IHC image processing procedure based on nucleus density, intensity of stained protein, and k-means clustering algorithm to develop an automated system for analyzing an IHC image. The proposed system can discriminate between normal and cancerous regions in an IHC image more effectively and display them visually. Furthermore, this system can automatically evaluate the stained protein expressions in the two regions which can help the pathologist to analyze the differential expressions of a marker protein from IHC images. Finally, we evaluated the proposed system on 150 real IHC liver cancer tissue images and compared the results with those obtained using support vector machine (SVM) and a previous work where the average of density of nuclei is used as the threshold to discriminate cancerous from normal regions.

Key-Words: - Immunohistochemistry, Tissue Image, Nuclei Segmentation, K-means Clustering, Support Vector Machine

1 Introduction

It was generally believed that the formation of cancer is due to activation of oncogenes or in activation of tumor suppressor genes. Many studies of genes associated with cancer have been conducted over the years, expected to find genes that cause cancer and genes that can inhibit cancer [1]. Several approaches have been developed to help diagnose cancer in the past. These approaches may include the use of ultrasound, magnetic resonance, and computerized topography scan, etc. However, the diagnosis of cancer should eventually be confirmed subject to the biopsy.

With the advances of microarray technology, microarray technology can be used to determine molecular markers and the type of cancer [2]. However, microarray gene expressions are usually only able to provide early stage information of gene

* An earlier version of this paper was presented at The Conference on Technologies and Applications of Artificial Intelligence (Domestic Track), National Chiao Tung University, Hsinchu, Taiwan, Nov. 19, 2010.

expressions. The differential gene expressions are still needed to be validated by quantified polymerase chain reaction (PCR). A complementary imaging technique to quantify protein activity is based on immunohistochemistry (IHC) images [3]-[9]. While a number of automated or semi-automated techniques are increasing in popularity, manual intervention is still necessary in order to resolve particularly challenging or ambiguous cases. This calls for new computer-aided tools to assist the pathologist in the examinations of their work. For this, an extensive review for automated analysis of IHC images has been made by Theodosiou et al. [10]. IHC can be used to detect whether the target antigen exists in tissue sections or not in order to discriminate between the cancerous and normal regions in a cancer tissue specimen. It can help us to recognize the location and distribution of marker proteins in different regions of the specimen. However, if the boundaries of normal and cancerous regions are not clear, the analysis results will be extremely affected by inter-observer variability [11]. Moreover, it is time-consuming to extract information from very large datasets by manual analysis.

Fig. 1 shows an example of IHC image of liver cancer tissue. This kind of images will be taken as the target image to be considered in this paper. The dark blue parts of the image are nuclei. The density of nuclei can be used to distinguish between normal and cancerous regions in the tissue. The region with a higher density of nuclei will be recognized as cancerous tissue and the region with a lower density of nuclei will be recognized as normal tissue. However, because tissues may have some deformation when embedding and cells are not neatly arranged in the same plane, many nuclei may be seen to be connected visually from the biopsy. Besides, since the composition of human liver cell is a cell with multiple nuclei, it does not necessarily have a nuclear membrane between nuclei. Hence, counting the number of nuclei per unit region is not an effective way to find the density of nuclei. One way to deal with this problem is to discriminate nuclei region from non-nuclei region based on the color information of nucleus before calculating the density of nuclei [12]. Another factor that will affect the accuracy of discrimination is the threshold for the density of nuclei. In this paper, we will consider this issue based on the k-means clustering algorithm. The k-means clustering has been a well-known unsupervised learning algorithm [13]. Di Cataldo et al. [14] have also demonstrated that the unsupervised approach, namely k-means clustering, overcomes the accuracy of a theoretically superior supervised method such as Support Vector Machine (SVM) [15] based on the experimental results on a heterogeneous dataset of IHC lung cancer tissue images. We will propose an improved IHC image processing procedure based on nucleus density, stained protein expressions, and k-means clustering algorithm to develop an automated system for discriminating normal from cancerous regions in IHC liver cancer tissue images. Experimental results on 150 real IHC images demonstrate that the proposed system achieves better results as compared to those obtained using support vector machine and a previous work where the average of density of nuclei is used as the threshold to discriminate cancerous from normal regions [12].

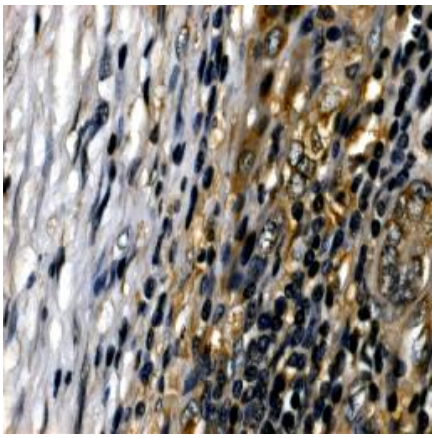

Fig. 1. An example of IHC image of liver cancer tissue.

2 Method

In this section, we introduce the methods and an improved automated system for IHC liver cancer image analysis. The work flow of the improved system is shown in Fig. 2. The work flow is mainly consisted of four steps as described below.

2.1 Recognizing nucleus and stained protein regions

The objective of the first step of IHC image analysis is to recognize nucleus and stained protein regions in the IHC liver cancer tissue image. The nuclei appear as dark blue color in the IHC image as shown in Fig. 1. Instead of using color deconvolution to deal with the effect of stains colocalization [9], we transform the original color IHC image acquired with a red-green-blue (RGB) camera to YIQ color space [16]. The reason to use YIQ color representation is that it takes advantage of human color-response characteristics. The Y component stands for the luma information (the

brightness in an image). The I and Q components stand for the chrominance information with color in the orange-blue range and purple-green range, respectively. We extract the I component of the image and enhance it with a threshold selection method [17]. Then, the B component of the original RGB color image is replaced by the enhanced I component of the converted YIQ color image. Since our goal here is to identify the regions of nucleus, we use the following formula to transform the image to gray-level image:

$$Y_i = 0.299R_i + 0.587G_i + 0.114B_i, i = 1, 2, \ldots, N \quad (1)$$

where Y_i denotes the gray level of pixel i and N the number of pixels of the image. R_i, G_i, and B_i represent the R, G, and B components of pixel i, respectively. This gray-level image is then converted to binary image. The segmented black color regions of the resulting binary image denote the nuclei. After recognizing the nucleus, we identify the region of stained protein in the original IHC image, where the brown stained regions indicate that the marker protein is present.

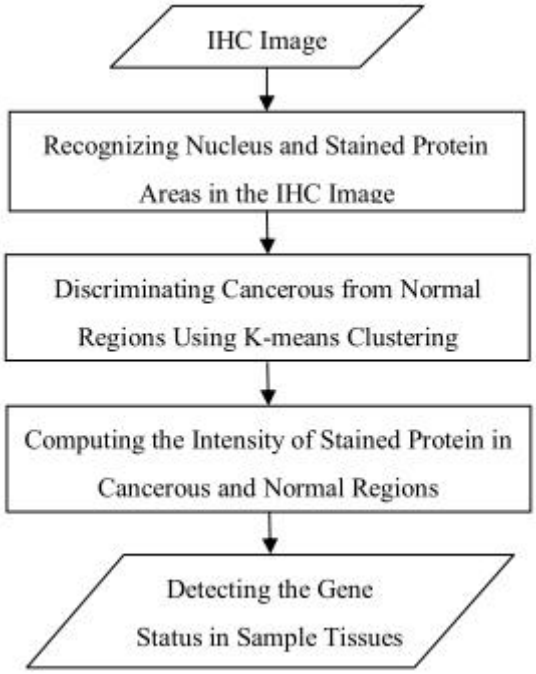

Fig. 2. Work flow of the proposed system for IHC liver cancer image analysis.

2.2 Discriminating cancerous from normal regions using k-means clustering

The purpose of this step is to distinguish between normal and cancerous tissue regions in the IHC image based on the information of nucleus density. The image is first segmented using a sliding window of size n x n, which can be defined by the user. The average nucleus density at segment i, DNi is calculated by the following formula [12]:

$$DN_i = \frac{C_i}{A_i}, i = 1, 2, \ldots, block\_number \quad (2)$$

where C_i is the number of pixels of nucleus regions at segment i of the image, A_i is the number of pixels of segment i, and *block_number* is the number of segments of the image. The average nucleus density at each segment is selected as the discriminant feature. After evaluating the average nucleus density of all segments, we apply the *k*-means clustering to separate the cancerous from normal image regions based on the discriminant feature. In statistics and machine learning, *k*-means clustering is an approach of cluster analysis which aims to separate observations into k clusters such that each observation is assigned to the cluster with the nearest mean; it is also referred to as Lloyd's algorithm [18]-[20].

Since the *k*-means clustering is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial data. Usually, the initial data are specified at random or by some heuristic data. In the proposed system, we will sort the nucleus density of all segments at first, and then the maximum density and minimum density are assigned to the initial data. This setting makes the clustering result more accurate in our analysis. Another problem with using *k*-means clustering algorithm is to determine the number of clusters, an inappropriate choice of k may yield poor results. The number of k clusters is set to 2 since only two classes, normal and cancerous tissue regions, are to be separated in this analysis.

2.3 Computing the intensity of stained protein in cancerous and normal regions

After discriminating cancerous from normal regions in an IHC tissue image, the third step is to compare the intensity of stained protein in both regions in order to detect the gene status in sample tissues. We first map the regions of stained protein obtained in Section 2.1 onto cancerous and normal regions of the IHC image. The average intensity of stained protein in cancerous region, denoted by *DP_CR*, is then evaluated by the following formula:

$$DP\_CR = \sum_{j=1}^{num\_cr} \sum_{(x,y)\in CA_j} f(x,y) \Big/ \sum_{j=1}^{num\_cr} (A_j - C_j) \quad (3)$$

where *num_cr* is the number of segments in cancerous region, A_j is the number of pixels of j^{th} segment in cancerous region, C_j is the number of pixels of nucleus region in j^{th} segment of cancerous region, $f(x, y)$ is the gray-level intensity at coordinate (x, y), and CA_j is the set of coordinates in j^{th} segment of cancerous region. Note that the average intensity of stained protein is calculated by excluding the region of nuclei in each segment. Similarly, the average intensity of stained protein in normal region, denoted by *DP_NR*, can be evaluated by the following formula:

$$DP\_NR = \sum_{j=1}^{num\_nr} \sum_{(x,y)\in NA_j} f(x,y) \Big/ \sum_{j=1}^{num\_nr} (B_j - N_j) \square \quad (4)$$

where *num_nr* is the number of segments in cancerous region, B_j is the number of pixels of j^{th} segment in normal region, N_j is the number of pixels of nucleus region in j^{th} segment of normal region, $f(x, y)$ is the gray-level intensity at coordinate (x, y), and NA_j is the set of coordinates in j^{th} segment of normal region.

2.4 Detecting the gene status in sample tissues

The final step of the proposed system is to detect the gene status in sample tissues based on the comparison of average intensity of stained protein in cancerous and normal regions obtained in Section 2.3. There are three gene statuses for the detection. If the average intensity of stained protein in cancerous region is greater than the one in normal region, the gene status will be recognized as "Over Expression". If the average intensity of stained protein in cancerous region is less than the one in normal region, the gene status will be recognized as "Under Expression". Otherwise, the gene status will be recognized as "no significant difference" which may imply the target marker protein should have no relationship with the target disease. In the case of "no significant difference", the proposed system also provides a user-defined threshold for distinguishing the average intensity of stained protein between cancerous and normal regions.

3 Performance measure

In this section, we introduce the performance measure used to evaluate the success of the proposed method.

The success of a test lies in its ability to retrieving correct results for a given task. In statistics, the *F*-score (also *F*-measure) is a measure for evaluating the performance of a test [21]. It considers both the precision *p* and recall *r* of the test to compute the score. The precision is defined as the fraction of tested results that are correct. The recall is defined as the fraction of correct results that are tested. In these terms, *F*-score is defined as the weighted harmonic mean of recall and precision. The formula used to compute the *F*-score here is as follows:

$$F = \frac{2rp}{r+p} \square \quad (5)$$

In the following section, we will use the *F*-score to evaluate the detecting performance of the proposed method and the method of previous work in [12].

4 Experimental results and discussion

The proposed system was implemented in Borland® C++ Builder, a well-known C++ development environment, and combined with the MATLAB®.

In this paper we have carried out an extensive experimental evaluation on 150 real IHC images [22] for three methods. The first method is the proposed system with *k*-means clustering (designated Method 1), the second one is SVM (designated Method 2) and the third one is a previous work [12] (designated Method 3) using the average of density of all nuclei as the threshold to discriminate cancerous from normal regions.

In Method 2, we first cut each IHC image into 64 blocks with each block size set to 32 x 32.Then we have a total of 9600 blocks of subimages. Among these subimages, we choose 1000 subimages with random for training, including 250 blocks of cancerous subimages and 750 blocks of normal subimages. The remaining 8600 subimages are used as the test dataset for classification.

The comparison of performance measures, average *F*-score and accuracy, are shown in Table 1. The average *F*-score obtained using Method 1, Method 2, and Method 3 is 82.49%, 56.77%, and 67.71%, respectively. It can be found that Method 1 overcomes the average F-score of Method 2 and Method 3 by 25.72% and 14.78%, respectively. On the other hand, Method 1 achieves an average accuracy of 89.84%, Method 2 achieves an average accuracy of 77.99%, and Method 3 achieves an average accuracy of 77.95%. Method 1 also overcomes the average accuracy of Method 2and Method 3 by 11.85% and 11.89%, respectively. The overall performance of Method 1 is better than that of Method 2 and Method 3 based on their average *F*-score and accuracy.

Table 1. Performance Comparisons of Method 1, Method 2, and Method 3.

Method	F-score (%)	Accuracy (%)
Method 1 (*k*-means)	**82.49**	**89.84**
Method 2 (SVM)	56.77	77.99
Method 3 (Average)	67.71	77.95

A snapshot of graphical user interface (GUI) of the proposed system is shown in Fig. 3. The images shown from left to right are original image, nucleus image, image with user-defined cancerous segments, and image with detected cancerous segments, respectively. The assumed cancerous regions in the image with user-defined cancerous segments can be specified by the human expert via the button "Setting Cancerous Segments" in the GUI. The size of sliding window for image segmentation is set to 32 x 32 in our experiments. The performance for the detection of cancerous segments can be evaluated by clicking the button "*F*-score" (see Section IV for details about the *F*-score). A snapshot of an illustrative example obtained with Method 1 on one of the real IHC image is shown in Fig. 4. It can be seen from Fig. 4 that Method 1 achieved an *F*-score of 0.8 (or 80%) and obtained a result of "Over Expression" for the estimated gene status.

5 Conclusion

We presented an improved computer-aided analysis system of IHC liver cancer tissue images. The procedure was mainly consisted of four steps, namely recognizing nucleus and stained protein regions, discriminating cancerous from normal regions using *k*-means clustering, computing the intensity of stained protein in cancerous and normal regions, and detecting the gene status in sample tissues. The proposed system was tested on 150 real IHC liver cancer tissue images. Experimental results demonstrated the high average *F*-score and accuracy achievable by the proposed system compared to the method of SVM and a previous work where the average of density of nuclei is used as the threshold to discriminate cancerous from normal regions.

Acknowledgements

This work was supported in part from the National Science Council, Taiwan [grant numbers: NSC 99-2221-E-024-010 and NSC 101-2221-E-024-024].

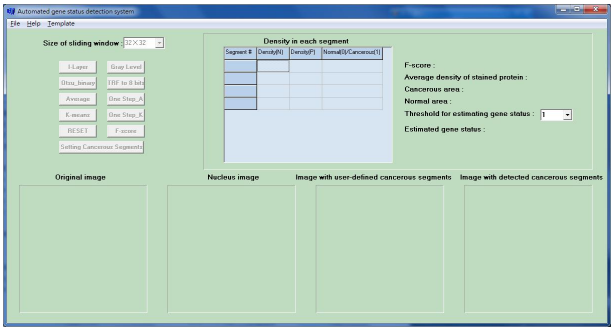

Fig. 3. A snapshot of graphical user interface (GUI) of the proposed system.

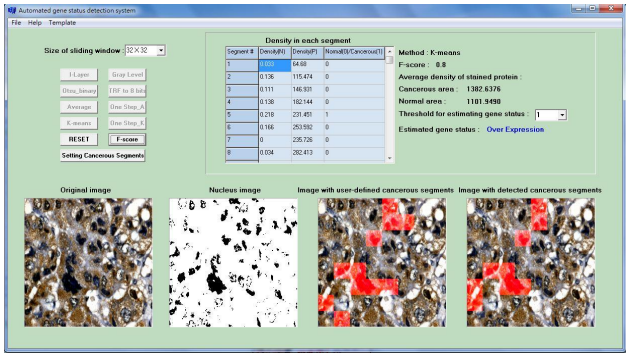

Fig. 4. A snapshot of experimental results obtained with Method 1 on one of the real IHC image.

References:

[1] S. Veerla, Cancer-Related Gene Regulation Mechanisms: Cancer Genetics, Transcription Factors, Gene Regulation Mechanisms, Bioinformatics, VDM Verlag, 2009.

[2] J.A. Warrington, R. Todd, D. Wong, *Microarrays and cancer research*, Eaton Publishing Company, 2002.

[3] E.M. Breya, Z. Lalanic, C. Johnstona, M. Wongc, L. V. McIntireb, P. J. Duked, C. W. Patrick, Automated selection of DAB-labeled tissue for immunohistochemical quantification, *Journal of Histochemistry and Cytochemistry*, Vol.51, 2003, pp. 575-584.

[4] J. K. Choi, U. Yu, O. J. Yoo, S. Kim, Differential coexpression analysis using microarray data and its application to human cancer, *Bioinformatics*, Vol.21, 2005, pp. 4348-4355.

[5] S. Di Cataldo, E. Ficarra, A. Acquaviva, E. Macii, Achieving the way for automated segmentation of nuclei in cancer tissue images through morphology-based approach: A quantitative evaluation, *Computerized Medical Imaging and Graphics*, Vol.34, No.6, 2010, pp. 453–461.

[6] S. Di Cataldo, E. Ficarra, A. Acquaviva, E. Macii, Automated segmentation of tissue

images for computerized IHC analysis, *Computer Methods and Programs in Biomedicine*, Vol.100, No.1, 2010, pp. 1-15.

[7] E. Ficarra, E. Macii, L. Benini, G. De Micheli, Computer-aided evaluation of protein expression in pathological tissue images, in: *19th IEEE Symposium on Computer-Based Medical Systems*, 2006, pp. 413-418.

[8] J. T. Jrgensena, Pharmacodiagnostics and targeted therapies: a rational approach for individualizing medical anticancer therapy in breast cancer, *Oncologist*, Vol.12, 2007, pp. 397-405.

[9] A. C. Ruifrok, D.A. Johnston, Quantification of histochemical staining by color deconvolution, *Anal. Quant. Cytol. Histol.*, Vol.23, 2001, pp. 291-299.

[10] Z. Theodosiou, I. Kasampalidis, G. Livanos, M. Zervakis, I. Pitas, K. Lyroudia, Automated analysis of FISH and immunohistochemistry images: a review, *Cytometry, Part A*, Vol.71, 2007, pp. 439-450.

[11] M. Lacroix-Triki, S. Mathoulin-Pelissier, J. Ghnassia, G. Macgrogan, A. Vincent-Salomon, V. Brouste, M. Mathieu, P. Roger, F. Bibeau, J. Jacquemier, High inter-observer agreement in immunohistochemical evaluation of HER-2/neu expression in breast cancer: a multicentre GEFPICS study, *EJC*, Vol.42, 2006, pp. 2946-2953.

[12] C. Y. Hsu, R. M. Hu, R. M. Chen, J. W. Ou, J. P. Tsai, *IHCREAD: an automatic immune-histochemistry image analysis tool*, in: Biomedical Engineering: Health Care Systems, Technology and Techniques, Springer, New York, 2011.

[13] A. K. Jain, R. C. Dubes, *Algorithms for clustering data*, Prentice Hall, Englewood Cliffs, 1988.

[14] S. Di Cataldo, E. Ficarra, E. Macii, Automated discrimination of pathological regions in tissue images: unsupervised clustering vs. supervised SVM classification, Commun. Comput. *Inform. Sci.*, 2008, pp. 344-356.

[15] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, 1998.

[16] J. D. Foley, A. van Dam, S. K. Feiner, J. F. Hughes, *Computer Graphics: Principles and Practice*, Reading, MA: Addison-Wesley, 1990.

[17] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Trans. Systems, Man, and Cybernetics*, Vol.9, 1979, pp. 62-66.

[18] J. A. Hartigan, M. A. Wong, Algorithm AS 136: a k-means clustering algorithm, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, Vol.28, 1979, pp. 100-108.

[19] S. Lloyd, Least squares quantization in PCM, *IEEE Trans. Information Theory*, Vol.28, 1982, pp. 129-137.

[20] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: *5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1, 1967, pp. 281-297.

[21] C. J. van Rijsbergen, *Information Retrieval*, Butterworth, London, 1979.

[22] J. W. Lu, J. G. Chang, K T. Yeh, R. M. Chen, Jeffrey J. P. Tsai, W. W. Su, R. M. Hu, Increased expression of PRL-1 protein correlates with shortened patient survival in human hepatocellular carcinoma, *Clinical and Translational Oncology*, Vol.14, 2012, pp. 287-293.