Colon Cancer Detection Using Whole Slide Histopathological Images

Liping Jiao[1], Qi Chen[1], Shuyu Li[1] and Yan Xu[1]

[1] School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China

Abstract—***Introduction:*** **Colon cancer is one of the major reasons of cancer-related deaths, whereas both the traditional and the current methods are complex to conduct.** ***Materials and Methods:*** **This paper presents a simple and effective approach for colon cancer detection by classifying cancer and non-cancer colon images based on computer assisted diagnosis. A classifier is developed using SVM, which has excellent performance in practice. Eighteen simple features, including gray-scale mean, gray-scale variance and 16 texture features extracted by Gray-Level Co-occurrence Matrix (GLCM) method, are chosen as the feature set.** ***Results:*** **In order to evaluate the accuracy of our classification, we calculate precision, recall and F-measure of different classifiers produced by using different feature combinations. And 3-fold cross-validation is applied. Three indicators precision, recall and F-measure are used to describe the performance of our system. Experiment results show that: when all features are used, the mean value of precision, recall and F-measure are 96.67% 83.33% 89.51% respectively.** ***Discussion:*** **These results demonstrate the great advantage of the method on colonic histopathology images' classification. The simple and efficient method will have great contributions on colon cancer detection.**

Keywords— **Colon cancer, Feature extraction, GLCM, Computer-aided diagnosis, SVM**

I. Introduction

Colon cancer occurs in the large intestine (colon) or the rectum (end of the colon) [1]. According to American Cancer Society, colon cancer is one of the major reasons of cancer-related deaths in the US [2]. The key step for confirming the diagnosis of malignancy and guiding treatment is to stain colon biopsy by Hematoxylin and Eosin (H&E) technique [3]. Fig. 1 shows some examples of stain colon biopsy images from normal people and abnormal people.

The traditional method for cancer detection is very time-consuming and labor-intensive. In order to distinguish the colon images, pathologists need to accumulate a large amount of experience through observing the labeled histopathological images. Therefore, a considerable mass of resources and manual labors are wasted. Thus, better diagnostic accuracy and faster diagnosis speed are required.

In recent years, computer technologies draw a lot of attentions due to the advantages of computer, including computational power, speed and storage capability.

Inventing an automated system for cancer detection (prostate cancer [3], breast cancer [4], etc.) based on computer-aided diagnosis, has become researchers' focus. Machine learning [5] is one of the great applications of computer-aided technique, because of its ability of human-like learning, which will improve its algorithm by experience automatically. So far, there are various studies [6, 7] about colonic cancer diagnosis aided by computer technology. However, these systems are very complex relatively. This paper aims at developing a more objective computer-aided technique for simply and effectively detecting colon carcinoma by classifying cancer and non-cancer images based on SVM. Because of the simple and effective of SVM [3], this method is selected.

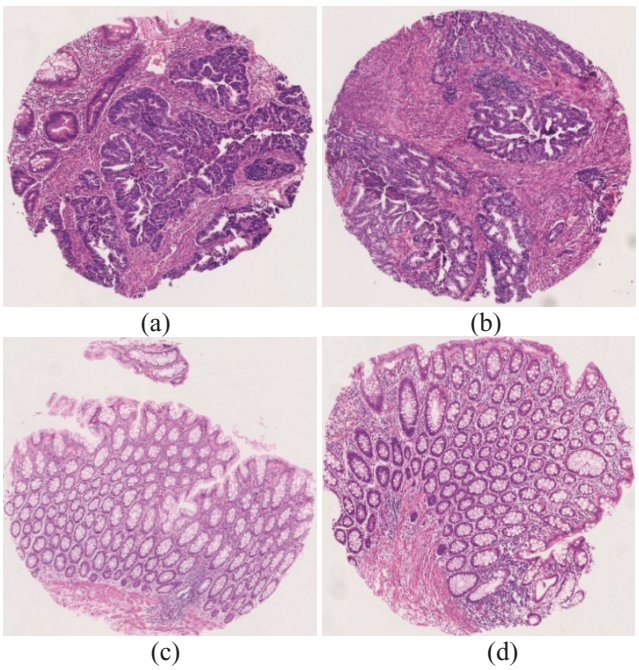

Fig 1: Examples of colon images
(a) & (b) Cancer images, (c) & (d) Non-cancer images

Feature extraction is a key step of building a classifier. Gabor filters, fractal dimension (FD), and multi-wavelet are commonly used in the diagnosis of prostate cancer [3], and Gray-Level Co-occurrence Matrix (GLCM) [8] transforms

M. Long (Ed.): World Congress on Medical Physics and Biomedical Engineering, IFMBE Proceedings 39, pp. 1283–1286, 2013.
www.springerlink.com

which introduced by Haralick et al. [9], is widely used in the diagnosis of colon cancer.

In this paper, depending on the analysis of images characteristics we combine mean, variance and other 16 features which are extracted by GLCM method to form a set of 18 features for classification. A 3-fold Cross-validation [10] pressure is applied to obtain the classification results. Precision, recall and F-measure [11] are calculated to characterize the result of our experiment. Finally, conclusions are drawn by analyzing the three indicators.

II. METHOD

SVM proposed in computer and statistics science is an excellent method of machine learning and has been extensively applied in classification and regression analysis.

In this section, a SVM classifier is produced based on colon image analysis and feature extraction.

A. Image Classification: SVM

SVM is a supervised learning algorithms based on recent advances in statistical learning theory. It is now established as one of the standard tool for machine learning due to its excellent performance in real-world applications, such as image classification, hand-written character recognition, text categorization, and so on.

The standard SVM takes a set of training and test data. Each training example is marked as belonging to one class, forming a label vector. SVM learns a linear decision to assign new examples into one or the other class according to the model built by training labeled training data, which consists of label vectors (positive and negative classes). The SVM model is a linear classification rule that can be used to discriminate test data into two classes.

B. Images Analysis

In order to extract useful features for classification, characteristics of images are analyzed here. Generally, as the H&E dye makes cancer images' color darker than the normal ones, so by observing the abnormal and normal colon tissue images (e.g. Fig. 1) we can come to conclusions as follows. Firstly, there is a sharp distinction between the grey levels of these two kinds of images. Secondly, compared to the normal images, abnormal ones have larger color ranges. Finally, significant texture differences can be observed in these two kinds of images.

C. Feature Extraction

a) Mean and Variance features

Mean (μ_f): pixel's gray values of ergodic nodules are added up and then divided by the nodes' area.

$$\mu_f = \frac{\sum_{(x,y\in R)} f(x,y)}{A} \tag{1}$$

In the above formula, μ_f represents mean value, and $f(x,y)$ stands for gray value of each point.

Variance (σ_f): also known as Grayscale consistency. It reflects the intensity of gray values' changing within the node areas.

$$\sigma_f = \sqrt{\frac{\sum_{(x,y\in R)} (f(x,y)-\mu_f)^2}{A}} \tag{2}$$

b) GLCM features

To extract the texture information of an image, Haralick proposes a second-order statistical method, named as Gray-Level Co-occurrence Matrix (GLCM) [6]. Here are some commonly used eigenvalues shown in (3), (4), (5) and (6).

Angular second moment (ASM): Always called energy, it is the sum of the GLCM elements' square, which indicates the distribution of gray level and texture coarseness.

$$\mathrm{ASM} = \sum_i \sum_j p(i,j)^2 \tag{3}$$

Contrast (CON): It is the GLCM's Contrast near the main diagonal moment of inertia. It reflects the clarity of the images and texture depth.

$$CON = \sum_i \sum_j (i-j)^2 p(i,j) \tag{4}$$

Correlation ($CORRLN$): It measures the similarity degree of GLCM elements in row or column directions, therefore the correlation values reflect the relevance of local gray images.

$$CORRLN = [\sum_i \sum_j ((i,j)p(i,j)) - \mu_x \mu_y / \sigma_x \sigma_y \tag{5}$$

Entropy (ENT): It reflects the randomness of image texture. When the GLCM's all values are equal, it obtains the maximum value. On the contrary, if the value is uneven, this value is small.

$$\mathrm{ENT} = -\sum_i \sum_j p(i,j) \log p(i,j) \tag{6}$$

The GLCM method for feature extraction is described as follows. Consider an image of size N×N pixels that has G kinds of Gray levels. Then, we can get a size G×G co-occurrence matrix $p_d(i,j)$ .Displacement vector is $d = d(dx,dy)$ [7]. Thus for the element (i,j) of p_d , where

gray levels are i and j respectively, the number of pixel pairs is determined by the formula (7).

$$p_d(i,j) = \left|\{((r,s),(t,v) : I(r,s) = i, I(t,v) = j)\}\right| \quad (7)$$

In the formula (r, s), (t, v) ∈N×N, (t, v) ∈(r+dx, s+dy), "| |" stands for element number which often called set potential.

A simple example is showed in the next tables. Table 1 is the original image, Table 2 is the image transformed by 45 degree ($d = d(1,1)$) orientations' GLCM, and Table 3 is the image transformed by 0 degree ($d = d(1,0)$) orientations' GLCM. So 16 texture features are obtained by using the four GLCM features mentioned above. Each of the four GLCM features contains four orientations (0 degree, 45 degree, 90 degree and 135 degree), so 16 features which represent the image texture features are obtained.

Table 1 4x4 original image

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

Table 2 GLCM d=(1,1)

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

Table 3 GLCM d=(1,0)

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

III. Experiments and Results

In order to test the classifier, we developed a platform hosting the algorithm mentioned above. Experiment is conducted on this platform to test the performance of classifier. Sixty whole slide histopathological images used in this paper including 30 normal and 30 abnormal images are provided by a hospital. These images are labeled by two experienced doctors and confirmed by a third one.

A 3-fold cross-validation is used during the experiment procedure to achieve the classification results based on an average strategy which is more accurate. Also, in order to verify the contribution of each feature, we repeated the classification process 18 times by adding a new feature to the last feature set to produce a new classifier and evaluate its performance by using evaluation parameters until all the 18 features are added into the feature sets. Table 4 provides the results of the above experiment.

Table 4 The performance of our method

Feature	Precision	Recall	F-measure
Mean	87.04%	80.00%	83.37%
+Variance	87.04%	80.00%	83.37%
+ASM(0degree)	85.40%	76.67%	80.80%
+ASM(45degree)	85.40%	76.67%	80.80%
+ASM(90degree)	84.80%	73.33%	78.65%
+ASM(135degree)	85.40%	76.67%	80.80%
+CON(0degree)	86.71%	73.33%	79.46%
+CON(45degree)	87.45%	73.33%	79.77%
+CON(90degree)	88.64%	70.00%	78.22%
+CON(135degree)	93.33%	73.33%	82.13%
+CORRLN(0degree)	96.30%	80.00%	87.40%
+CORRLN(45degree)	96.30%	80.00%	87.40%
+CORRLN(90degree)	96.30%	83.33%	89.35%
+CORRLN(135degree)	96.67%	83.33%	89.51%
+ENT(0degree)	96.67%	83.33%	89.51%
+ENT(45degree)	96.67%	83.33%	89.51%
+ENT(90degree)	96.67%	83.33%	89.51%
+ENT(135degree)	96.67%	83.33%	89.51%

Cross-Validation

Cross-Validation is a method for proving the performance of classifier by dividing data into two segments: one used as training model and the other used as test model.

In k-fold cross-validation whole data is first divided into k equal parts. Then, k-1 parts are randomly selected as training set for testing the rest samples. Finally, repeat this process for all parts and calculate the entire results' average value to evaluate the accuracy of our experiments.

Evaluation Parameters

To evaluate our experimental results, precision, recall and F-measure are calculated. Precision stands for the proportion of true cancer patients among those diagnosed as cancer patients. Recall represents the proportion of true cancer patients diagnosed as cancer patients among those true cancer patients. F-measure value presents the whole performance of the classifier by taking comprehensive consideration of precision and recall. In the following formulas, we assume that: TP stands for the number of abnormal images which diagnosed as cancer, FP represents the number of normal images which diagnosed as cancer, TN stands for the number of abnormal images which diagnosed as no

disease, and FN represents the number of normal images which diagnosed as no disease.

$$precision = \frac{TP}{TP + FP} \tag{8}$$

$$recall = \frac{TP}{TP + TN} \tag{9}$$

$$F - measure = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{10}$$

IV. Dissussion

In this paper, we combine different features to build a SVM classifier to detect colon cancer by classifying the colon images from cancer and non-cancer people. And we get the experiment results based on 3-fold cross-validation.

According to Table 4, effect of the classifying decreases due to some features, while increases when other features are added, so there's a certain correlation among different features. But in a general aspect, the tendency of the test result presents a rise. Thus, no feature is abandoned. As the final result turns out, the classifier using 18 features are of high quality for colon image classification.

The Fig 1 shows that normal cells are oval which only have a layer of cell membrane and arranged regular. In the future work, these characteristics will be considered to further improve the performance of the classifier.

V. Conclusion

The aim of this research is to develop an easier method for colon cancer diagnosis. We achieve this purpose by producing a simple classifier using SVM based on 18 simple features extracted from the whole slide histopathological images to classify cancer and non-cancer groups. The experimental results demonstrate that the approach proposed in this paper is very efficient and make a great contribution to colon cancer detection.

Acknowledgment

This work was supported by Grant 61073077 from National Science Foundation of China and Grant SKLSDE-2011ZX-13 from State Key Laboratory of Software Development Environment in Beihang University in China.

References

1. http://ehealthmd.com/content/what-colon-cancer.
2. http://www.nlm.nih.gov/medlineplus/ency/article/000262.htm.
3. Huang P W, Lee Cheng-Hsiung, Lin Phen-Lan (2009) Support vector classification for pathological prostate images based on texture features of Multi-Categories. IEEE, International Conference on Systems, Man, and Cybernetics, San Antonio, TX, USA, 2009, pp 912–916.
4. Mu T T, Asoke K-N (2007) Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM–RBF classifier. Journal of the Franklin Institute 344 (3-4):285-311.
5. Bradley L Whitehall, Stephen C-Y Lu (1991) Machine learning in engineering automation—The present and the future. Computers in Industry 17(2-3):91-100.
6. Konga J, Sertela O, Shimada H et al. (2009) Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. Pattern Recognition 42:1080-1092.
7. Atlamazoglu V, Yova D, Kavantzas NI et al. (2001) Texture analysis of fluorescence microscopic images of colonic tissue sections. IFMBE Proc. vol. 39, World Congress on Med. & Biomed. Eng. & Comput, 2001, pp 145–151.
8. Sieler A, Tanougast C, Bouridane A. (2010) A scalable and embedded FPGA architecture for efficient computation of grey level co-occurrence matrices and Haralick textures features. Microprocessors and Microsystems 34 (1):14-24.
9. Haralick R M, Shanmugam K., Dinstein I. (1973) Textural Features for Image Classification. IEEE Proc. Trans. on Systems, Man and Cybernetics, 1973, pp 610-621.
10. Fukunaga K (1990) Introduction to Statistical Pattern Recognition. Academic, New York, 1990.
11. Egghe L. (2007) Existence theorem of the quadruple (P, R, F, M): Precision, recall, fallout and miss. Information Processing & Management 43(1):265-272.

Author: Yan Xu
Institute: School of Biological Science and Medical Engineering, Beihang University
Street: No.37 XueYuan Road, HaiDian District
City: Beijing 100191
Country: China
Email: xuyan@buaa.edu.cn