Pattern Analysis for detecting Pathology Using Haralick Texture Features

S. Simonthomas
Department of Computer Science and Engineering, A.C.T College of Engineering and Technology, Nelvoy, Chengalpattu.

Abstract
Cancer is a diseases that characterized by out of control cell growth. There are different types of cancer disease, and each is classified by the type of cell that is initially affected by the disease. Cancer harms the body when damaged cells divide uncontrollably to form lumps or masses of tissue called tumours. In this paper haralick texture features used to detect the cancer disease by using a segmented images of cell tissues. it characterize the biomedical images in pattern recognization a tissue image and this tissue quantification based on the pattern classifications. then the given tissue image satisfies the haralick texture means it is a normal tissue. the calculation is based on pixel values in an image. The accuracy has been calculated based on SVM classifier. Working with the tissue images, our experiments gives that the proposed a 95% accuracy compare to the existing one. and the accuracies for the tissue image quantification.

Keywords
Cancer diagnosis, GLCM, Haralick Texture features, SVM, structural pattern recognition.

1. Introduction

Digital pathology systems are becoming increasingly important due to the increase in the amount of digitalized biopsy images and the need for obtaining objective and quick measurements. The implementation of these systems typically requires a deep analysis of biological deformations from a normal to a cancerous tissue as well as the development of accurate models that quantify the deformations. These deformations are typically observed in the distribution of the cells from which cancer originates, and thus, in the biological structures that are formed of these cells. For example, colon adenocarcinoma, which accounts for 90%–95% of all colorectal cancers, originates from epithelial cells and leads to deformations in the morphology and composition of gland structures formed of the epithelial cells[1]. Moreover, the degree of the deformations in these structures is an indicator of the cancer malignancy (grade). Thus, the correct identification of the deformations and their accurate quantification are quite critical for precise modeling of cancer.

Although digital pathology systems are implemented for different purposes, including segmentation[2] and retrieval [3], most of the research efforts have been dedicated to tissue image classification. Conventional classification systems are typically based on the use of statistical pattern recognition techniques [4], in which -dimensional feature vectors are extracted to model tissue deformations (i.e., deformations observed in a tissue structure as a consequence of cancer). Different approaches have been proposed for feature extraction. The first group directly uses pixel-level information. These approaches employ first order statistics of pixels [5], [6] and/or higher order texture models such as cooccurrence matrices [7], local binary patterns [8], multi-wavelets [9], and fractals [10]. While textural features can usually model the texture of small regions in an image well, they lose the information of local structures, such as the organization characteristics of glands, when they model the entire tissue image.

The second group relies on the use of component-level information. These approaches extract their features by identifying histological components, mostly cells, within a tissue and quantifying their morphology and organization characteristics. Morphological approaches quantify the shape and size characteristics of the components[4]. Similar to the textural approaches, they do not consider the organization of the components in feature extraction. In order to model the organization characteristics, it has been proposed to use structural approaches that quantify an image by constructing a graph on its tissue components and extracting global features on the constructed graph[13]. However, since these global features are extracted over the entire graph, they cannot adequately model the organization characteristics of local structures such as glands in a tissue.

More importantly, the extraction of graph features, which are in the form of -dimensional vectors, is only an approximation to the entire graph structure and cannot retain all the relations among the graph nodes (tissue components) [7]. On the other hand, structural pattern recognition techniques allow using graphs directly, without representing them by fixed size d-dimensional feature vectors. Although such techniques are used in different domains including image segmentation [9], object categorization [2], and protein modeling [1], they have not been considered by the previous structural approaches for tissue image classification. Moreover, these previous approaches have not represented an image

Manuscript received November 5, 2015
Manuscript revised November 20, 2015

as a set of sub graphs that are similar to the query graphs and used them in classification.

2. Dataset

We test our model on 3236 microscopic images of colon tissues of 258 patients. Tissues are stained with hematoxylin-andeosin. Images are digitized at microscope objective lens and pixel resolution is 640 480. Note that after constructing a graph, the magnification and resolution only affect the window size. For larger magnifications and/or resolutions, the window size should be selected large enough so that windows capture sufficient information from the key regions. Each image is labeled with one of the three classes: normal, low-grade cancerous, and high-grade cancerous. The images are randomly divided into training and test sets.

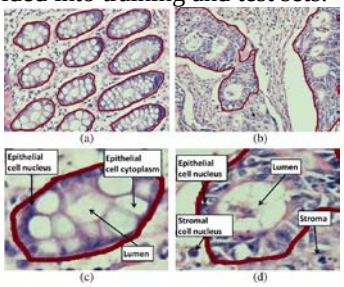

Fig. 1: Colon adenocarcinoma changes the morphology and composition of colon glands. This figure shows the gland boundaries on (a) normal and (b) cancerous tissue images. It also shows the histological tissue components the text will refer to on (c) normal and (d) cancerous gland images.

The training set contains 1644 images of 129 patients. It includes 510 normal, 859 low-grade cancerous, and 275 high-grade cancerous tissue images. The test set contains 1592 images of the remaining 129 patients. It includes 491 normal, 844 low-grade cancerous, and 257 high-grade cancerous tissue images. the Cancer Reform Strategy (2007) and review of the current data sets, the Data Sets Service is working with the National Cancer Intelligence Network (NCIN) to revise the data sets. The Cancer Genome Atlas, the Broad Genome Data Analysis Centre designs and operates scientific data and analysis pipelines which pump terabyte-scale genomic datasets through scores of quantitative algorithms, in the hope of accelerating the understanding of cancer.

3. Methodology

This paper presents a new approach to tissue image classification. This approach models a tissue image by constructing an attributed graph on its tissue components and describes what a normal gland is by defining a set of Haralick Texture Features. We propose a method based on the creation of a new image modality consisting in a grayscale map where the value of each pixel indicates its probability of belonging to a cell nuclei. This probability map is calculated from texture and scale information in addition to simple pixel color intensities. The resulting modality has a strong object-background contrast and evens out the irregularities within the nuclei or the background and feature extraction from the key regions with support vector classifier. this may cause misleading results due to the existence of nongland regions in the image, which are irrelevant in the context of colon adenocarcinoma diagnosis.

A. Gray Level Co-occurrence Matrix:

A GLCM P[i,j] is defined by specifying displacement vector d=(dx,dy). Counting all pairs of pixels separated by d, having gray levels i and j. GLCM Measures are, Entropy-Randomness of gray level distribution, Energy-uniformity of gray level in a region, Contrast-Measure of difference between gray levels and Homogeneity-Measure of similarity of texture[12]. Suppose that the gray tone appearing in each resolution cell is quantized' to Ng levels. Let Lx = {1,2, ,Nx} be the horizontal spatial domain, Ly = {1,2,. ,NY} be the vertical spatial domain, and G = {1,2, .*,NgJ be the set of Ng quantized gray tones. The set Ly x Lx is the set of resolution cells of the image ordered by their row-column designations. The image I can be represented as a function which assigns some gray tone in G to each resolution cell or pair of coordinates in Ly x Lx; I: Ly x Lx→G.
These measures are arrays termed angular nearest-neighbor gray-tone spatial-dependence matrices, and to describe these arrays we must emphasize our notion of adjacent or nearest-neighbor resolution cells themselves. We consider a resolution cell-excluding those on the periphery of an image, etc.-to have eight nearest-neighbor resolution[14].

B. Haralick Texture Features:

Segmentation partitions an image into distinct regions containing each pixels with similar attributes. To be meaningful and useful for image analysis and interpretation, the regions should strongly relate to depicted objects or features of interest. Meaningful segmentation is the first step from low-level image processing transforming a greyscale or colour image into one or more other images to high-level image description in terms of features[4], objects, and scenes. The success of image analysis depends on reliability of segmentation, but an accurate partitioning of an image is generally a very challenging problem. Haralick's texture features were calculated using the kharalick() function. The basis for these features is the gray-level co-occurrence matrix. This matrix is square with dimension Ng, where Ng is the

number of gray levels in the image. Element [i,j] of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j [6].

C. SVM Classifier:

Classifying data has been one of the major parts in machine learning. The idea of support vector machine is to create a hyper plane in between data sets to indicate which class it belongs to. The challenge is to train the machine to understand structure from data and mapping with the right class label, for the best result, the hyper plane has the largest distance to the nearest training data points of any class. classifies each row of the data in Sample, a matrix of data, using the information in a support vector machine classifier structure SVMStruct, created using the svmtrain function. Like the training data used to create SVMStruct [16], Sample is a matrix where each row corresponds to an observation or replicate, and each column corresponds to a feature or variable. Therefore, Sample must have the same number of columns as the training data. This is because the number of columns defines the number of features. Groupindicates the group to which each row of Sample has been assigned[17].

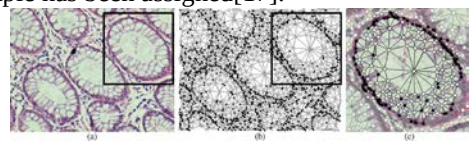

Fig. 2: An illustration of the graph generation step: (a) an example normal tissue image, (b) the tissue graph generated for this image, and (c) a query graph generated to represent a normal gland.

This approach models a tissue image by constructing features for haralick texture on its tissue components and describes what a normal tissue image is by defining a set of smaller query classification[1]. It searches the query thet classifies, which correspond to nondeformed normal tissue images, over the entire tissue graph to locate the attributed subgraphs that are most likely to belong to a normal image structure[9].

1. Angular second moment:

The angular second moment gives a strong measure of uniformity and can be defined as

$$f1 = \sum_i \sum_j P^2\ i,j$$

2. Contrast:

Contrast is defined as,

$$f2 = \sum_{n=0}^{N_g} n^2$$

where, $n = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P_{i,j} . |i-j$

Ng = number of gray levels

3. Correlation:

The correlation feature is a measure of gray-level linear dependency of the image.

$$f3 = \frac{\sum_i \sum_j P_{i,j} - \mu^2}{\sigma^2}$$

where μ and σ are the mean deviation and standard deviation of the co-occurrence matrix respectively.

Mean: It is a measure of brightness,

$$\mu_P = \frac{1}{n^2} \sum_{r=0}^{n-1} \sum_{s=0}^{n-1} P_{r,s}$$

$P_{r,s}$ is pixel at location(r,s).

Standard deviation: It is a measure of contrast,

$$\sigma_P = \left[\frac{1}{n^2} \sum_{r=0}^{n-1} \sum_{s=0}^{n-1} [P_{r,s} - \mu_P]^2 \right]^{1/2}$$

4. Sum of squares:Variance

Variance is the measure that tells us by how much the gray levels are varying from the mean.

$$f4 = \sum_i \sum_j (i,j) P_{i,j} - \mu^2$$

5. Inverse Difference Moment:

Inverse difference moment is the measure of local homogeneity and is defined as

$$f5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} P(i,j)$$

Homogeneity: Homogeneity is the measure that increases with less contrast in the window.

$$\text{Homogeneity} = \sum_i \sum_j \frac{P_{i,j}}{1+|i-j|}$$

6. Sum Average:

$$f6 = \sum_{i=2}^{2N_g} i\ P_{x+y}(i)$$

7. Sum Variance:

$$f7 = \sum_{i=2}^{2N_g} (i - f8)^2\ P_{x+y}(i)$$

8. Sum Entropy:2

The entropy of a sum of independent random vectors.

$$f8 = -\sum_{i=2}^{2N_g} P_{x+y}(i) \log\{P_{x+y}(i)\}$$

9. Entropy:

It is a measure of randomness,

$$f_9 = -\sum_{b=0}^{L-1} P(b) \log_2 P(b)$$

where, p(b) N(b)/n^2 for $\{0 \le b \le L-1\}$
where L is the number of different values which pixels can adopt, N(b) = number of pixels of amplitude (b) in the pixel window of size 'n×n'.

10. Difference Variance:

variance measures how far a set of numbers is spread out.

f_{10} = variance of P_{x-y}

11. Difference Entropy:

$$f_{11} = -\sum_{i=0}^{N_g-1} P_{x-y}(i) \log\{P_{x-y}(i)\}$$

12,13. Information of Correlation:

$$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}}$$

$$f_{13} = (1 - \exp[-2.0(HXY2 - HXY)])^{1/2}$$

$$HXY = -\sum_i \sum_j P(i,j) \log(P(i,j))$$

Where *HX* and *HY* are entropies of p_x and p_y , and

$$HXY\,1 = -\sum_i \sum_j P(i,j) \log\{P_x(i)P_y(j)\}$$

$$HXY\,2 = -\sum_i \sum_j P_x(i)P_y(j) \log\{P_x(i)P_y(j)\}$$

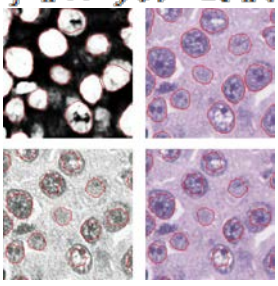

Fig. 3: Segments the nucleolus potion and extract the features based on Haralick texture features.

In pre-processing the given color image is converts into grayscale image for easy calculation of pixels values. And segmentation that partitions the image into multiple segments, haralick textures extracts the features based on the given tissue image. It shows an example of an incorrectly classified image together with the three-most similar subgraphs identified for each query graph.As shown in this figure, some of the identified in pixels values. The graph shows that the use of the median and the most dissimilar matching scores in sub graph selection decreases the accuracy of the hybrid model, which uses the most similar matching scores.

14.Maximal Correlation Coefficient:

Correlation coefficient, indicates the strength and direction of a linear relationship between two random variables.

$$f_{14} = (\text{Second largest eigenvalues of } Q)^{1/2}$$

Where,

$$Q(i,j) = \sum_k \frac{P(i,k)\,P(j,k)}{P_x(i)P_y(k)}$$

The text for Chosen distance d we have four angular gray-tone spatial-dependency matrices. Hence we obtain a set of four values for each of preceding 14 measures.

IV. Experimental Result

We report the test set accuracies obtained by the haralick texture features for the samples of each class, also leading to the highest overall accuracy. We also provide the confusion matrix for each tissue images. Almost all comparison algorithms extract their features on entire tissue pixels of an image. Moreover, in some images, these irrelevant regions may be larger than gland regions, and hence, they contribute to the extracted features more than the gland regions. The proposed model using haralick texture features classifies the accuracy of cancer disease by using learning support vector machine.

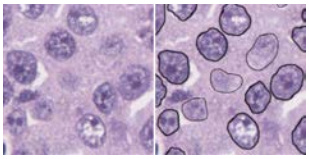

Fig. 4: magnified 250 × 250 region of a frame and (right) the same region with the nuclei delineated by a pathologist. Nuclei delineated with a thinner outline are hard to distinguish from the background. Dark and bright areas can indiscriminately occur inside and outside nuclei.

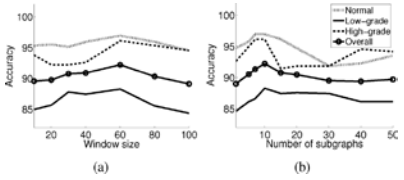

Fig. 5: Test set accuracies as a function of the external model parameters: (a)the window size W and (b) the number N of selected subgraphs.

We attribute this accuracy decrease to the following: In the subgraphs with the most dissimilar matching scores, the identified key regions mostly correspond to nongland

regions. In the subgraphs with the median matching scores, the key regions correspond to either gland or nongland regions. Since nongland regions exist in all classes, the effectiveness of the extracted features decreases in differentiating the images of different classes.

V. Conclusion and Future Work

This work presents a Haralick texture features that makes use of Segmentation with pattern recognition techniques for tissue image classification. This model proposes to represent a tissue image as an attributed graph of its components and characterize the image with the properties of its various features. The main contribution of this work is on the localization and characterization of the haralick features. The proposed model uses inexact pattern recognization. The proposed hybrid model is tested on 3236 colon tissue images. The experiments provides better results for detecting the cancer disease, reveal that this model gives significantly more accurate results compared to the existing approaches, which use only statistical pattern recognition techniques to quantify the tissue deformations. One future work is to develop a better approximation algorithm for graph edit distance calculations. For example, this algorithm may also consider the relations among the decomposed sub graphs. Moreover, one can modify the use of the BFS algorithm, or devise a new algorithm, in query graph and subgraph construction for more flexible matches.

REFERENCES

[1] R. Haralick, K. Shanmugam, and I. Dinstein, (1973) "Textural Features for Image Classification", IEEE Trans. on Systems, Man and Cybernetics, SMC–3(6):610–621

[2] Sumeet Dua, Senior Member, IEEE, U. Rajendra Acharya, Pradeep Chowriappa "Wavelet-Based Energy Features for Glaucomatous Image Classification" VOL. 16, NO. 1, JANUARY 2012

[3] U.Rajendra Acharya, Sumeet Due, Xian Du,and Vinitha Sree S "Automated Diagnosis of Glaucoma Using Textural and Higher Order Spectra Features"

[4] F. Yu and H. H. S. Ip, "Semantic content analysis and annotation of histological image," Comput. Biol. Med., vol. 38, no. 6, pp. 635–649, 2008.

[5] Bino Sebastian V, A. Unnikrishnan and Kannan Balakrishnan "grey level co-occurrence matrices: generalisation and some new features" (IJCSEIT), Vol.2, No.2, April 2012

[6] F. I. Alam, R. U. Faruqui, (2011) "Optimized Calculations of Haralick Texture Features", European Journal of Scientific Research, Vol. 50 No. 4, pp. 543-553

[7] Brachtel E, Yagi Y: Digital imaging in pathology–current applications and challenges. J Biophotonics 2012, 5:327-35.

[8] Hamilton PW, Wang Y, McCullough SJ: Virtual microscopy and digital pathology in training and education. APMIS 2012, 120:305-15.

[9] Daniel C, Rojo MG, Klossa J, Della Mea V, Booker D, Beckwith BA, Schrader T: Standardizing the use of whole slide images in digital pathology. Comput Med Imaging Graph 2011, 35:496-505.

[10] A. Tabesh, M. Teverovskiy, H. Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," IEEE Trans.Med. Imag., vol. 26, no. 10, pp. 1366–1378, Oct. 2007.