

LOCAL OBJECT PATTERNS FOR TISSUE IMAGE REPRESENTATION AND CANCER CLASSIFICATION

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Gülden Olgun
July, 2013

I certify that I have read this thesis and that in my opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. ıgdem Gündüz Demir(Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Hakan Ferhatosmanoğlu

I certify that I have read this thesis and that in my opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Süleyman Serdar Kozat

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

LOCAL OBJECT PATTERNS FOR TISSUE IMAGE REPRESENTATION AND CANCER CLASSIFICATION

Gülden Olgun

M.S. in Computer Engineering

Supervisor: Assist. Prof. Dr. Çiğdem Gündüz Demir

July, 2013

Histopathological examination of a tissue is the routine practice for diagnosis and grading of cancer. However, this examination is subjective since it requires visual interpretation of a pathologist, which mainly depends on his/her experience and expertise. In order to minimize the subjectivity level, it has been proposed to use automated cancer diagnosis and grading systems that represent a tissue image with quantitative features and use these features for classifying and grading the tissue. In this thesis, we present a new approach for effective representation and classification of histopathological tissue images. In this approach, we propose to decompose a tissue image into its histological components and introduce a set of new texture descriptors, which we call local object patterns, on these components to model their composition within a tissue. We define these descriptors using the idea of local binary patterns. However, we define our local object pattern descriptors at the component-level to quantify a component, as opposed to pixel-level local binary patterns, which quantify a pixel by constructing a binary string based on relative intensities of its neighbors. To this end, we specify neighborhoods with different locality ranges and encode spatial arrangements of the components within the specified local neighborhoods by generating strings. We then extract our texture descriptors from these strings to characterize histological components and construct the bag-of-words representation of an image from the characterized components. In this thesis, we use two approaches for the selection of the components: The first approach uses all components to construct a bag-of-words representation whereas the second one uses graph walking to select multiple subsets of the components and constructs multiple bag-of-words representations from these subsets. Working with microscopic images of histopathological colon tissues, our experiments show that the proposed component-level texture descriptors lead to higher classification accuracies than the previous textural approaches.

Keywords: Digital pathology, tissue image representation, classification, texture, local patterns, graph walks, colon cancer.

ÖZET

KANSER SINIFLANDIRMA VE DOKU GÖRÜNTÜ TEMSİLİNDE LOKAL NESNE DESENLERİ

Güliden Olgun

Bilgisayar Mühendisliği, Yüksek Lisans

Tez Yöneticisi: Assist. Prof. Dr. Çiğdem Gündüz Demir

Temmuz, 2013

Histopatolojik doku incelemesi kanser teşhis ve derecelendirmesinde rutin olarak uygulanan yöntemdir. Fakat, bu inceleme pataloğun uzmanlığına ve deneyimine bağlı olan görsel çıkarımlar gerektirdiği için öznellik içerir. Sonuçlardaki özneliliğin etkisini azaltmak için doku görüntüsünü nicel özelliklerle temsil eden ve bu özellikleri kullanarak doku sınıflandırması ve derecelendirmesi yapan otomatik kanser tanı ve derecelendirme sistemleri önerilmiştir. Bu tezde, etkili bir şekilde histopatolojik doku görüntülerini temsil etmek ve sınıflandırmak için yeni bir yaklaşım sunulmuştur. Bu yaklaşımda, doku görüntülerinin histolojik bileşenlerine ayrılması önerilmiş ve bu bileşenlerin doku içerisindeki dağılımını modellemek için lokal nesne desenleri olarak adlandırdığımız yeni bir grup örgüsel tanımlayıcı ortaya konulmuştur. Bu tanımlayıcılar, lokal ikili desenler yönteminin mantığı kullanılarak tanımlanmıştır. Ancak, pikseli, komşularının göreceli yoğunluğuna göre ikili bir dizi kurarak niceleyen piksel seviyesindeki lokal ikili desenlerin aksine, doku bileşenlerini nicelemek amacıyla, lokal nesne desen tanımlayıcıları bileşen seviyesinde tanımlanmıştır. Bu amaçla, değişik lokallik alanındaki komşuluklar belirlenmiş ve belirlenen komşuluklardaki bileşenlerin uzaydaki düzeni kodlanmıştır. Sonrasında, histolojik bileşenleri karakterize etmek amacıyla, bu dizilerden örgüsel tanımlayıcılar çıkartılmış ve bu şekilde karakterize edilmiş bileşenlerden resmin kelime - torbası temsili oluşturulmuştur. Bu tezde, bileşenlerin seçilmesi için iki yaklaşım kullanılmıştır: İlk yaklaşım kelime - torbası temsili çıkartmak için tüm bileşenleri kullanırken, ikinci yaklaşım çizge yürümesi ile birden fazla bileşen alt kümesi seçmiş ve bunlardan birden fazla kelime - torbası temsili oluşturmuştur. Mikroskopik histopatolojik kolon doku görüntüleri üzerinde yaptığımız deneyler, önerilen bileşen seviyesindeki örgüsel tanımlayıcıların önceki örgüsel yaklaşımlara göre daha yüksek doğruluk oranları verdiğini göstermektedir.

Anahtar sözcükler: Dijital patoloji, doku görüntü temsili, sınıflandırma, örgü, lokal desenler, izge yürümesi, kolon kanseri.

Acknowledgement

I would like to thank my advisor Assist. Prof. Dr. ıdem Gndz Demir for her endless support. She always trust me even I am in predicament and she gives all her efforts to help me and to teach me. To be honest, without her help and guidance, I would not finish this thesis. I am glad to work with Assist. Prof. Dr. ıdem Gndz Demir and for being a student of her. I also thank Assoc. Prof. Dr. Hakan Ferhatosmanoėlu and Assist. Prof. Dr. Sleyman Serdar Kozat for spending their precious time for my thesis.

I want to thank my dear parents Nurcan and Naci Olgun and and my beloved one,my elder sister, Gizem Olgun for their support, patience, love and for everything. Also, I thank my dear friends Gizem Mısırlı, Meryem Mudara, Pınar Karaoėlan, Seda Dumlu, Pren Gler, zlem Gr, Glbanu Altunok, Berk Glmezoėlu and Nurulhude Baykal for their motivation, love and support. Moreover, I thank my dearest friend İsmail Sengr Altıngvde for his motivation. Finally, I want to thank my research group members Can Fahrettin Koyuncu, Salim Arslan and Tun Gltekin.

Contents

1	Introduction	1
1.1	Contribution	2
1.2	Outline of the Thesis	4
2	Background	5
2.1	Domain Description	5
2.2	Automated Cancer Classification	7
2.2.1	Textural Methods	7
2.2.2	Structural Methods	14
3	Methodology	18
3.1	Tissue Image Decomposition	18
3.2	Local Object Patterns	20
3.3	Tissue Classification	23
4	Experimental Results	25

4.1	Dataset	25
4.2	Comparisons	26
4.2.1	Textural Approaches	26
4.2.2	Structural Methods	28
4.3	Parameter Selection	30
4.3.1	Simple Approach	30
4.3.2	GraphWalk Approach	32
4.3.3	Comparison Algorithms	32
4.4	Results	33
4.5	Parameter Analysis	36
4.5.1	Simple Approach	36
4.5.2	GraphWalk Approach	39
5	Conclusion	42

List of Figures

2.1	An image of a normal colon tissue stained with hematoxylin and eosin, which is the routinely used technique to stain biopsies in hospitals.	6
2.2	Examples of colon tissue images: (a)-(b) normal tissues, (c)-(d) low-grade cancerous tissues, and (e)-(f) high-grade cancerous tissues.	8
2.3	The Voronoi diagram for randomly selected 20 points	15
2.4	The Delaunay triangulation for the 20 random selected points	16
2.5	The Gabriel's graph for the same randomly selected 20 points	16
2.6	Minimum spanning tree constructed for the same randomly selected 20 points	17
3.1	Examples of tissue images and their located objects. In these images, (a) a normal tissue image, (c) a low-grade cancerous tissue image, (e) a high-grade cancerous tissue image and (b)(d)(f) the objects located on these tissue images. Here purple, pink, and white objects are represented with purple, pink, and cyan, respectively.	19

3.2	Extracting local object patterns for the objects shown in black. Here m is selected as 4, thus $S = \bigcup_{j=0}^m 2^j$ -LOP. Sixteen-nearest neighbors of the selected objects are indicated on the examples with their orders.	22
4.1	Test set accuracy as a function of the minimum circle radius r_{min}	37
4.2	Test set accuracy as a function of the highest degree m	38
4.3	Test set accuracy as a function of the cluster number k	39
4.4	Test set accuracy as a function of the subset number N	40
4.5	Test set accuracy as a function of the visited object number L . .	41

List of Tables

2.1	The most commonly used intensity histogram features	9
2.2	Cooccurrence matrix features	11
2.3	The mostly used run-length matrix features	12
4.1	For textural approaches, considered parameter values are listed. The parameter values selected by three fold cross-validation are indicated as bold.	31
4.2	For structural approaches, considered parameter values are listed. The parameter values selected by three fold cross-validation are indicated as bold.	31
4.3	Test set results obtained by our proposed algorithms <i>Simple</i> ap- proach and <i>GraphWalk</i> Approach and the textural comparison al- gorithms	34
4.4	Test set results obtained by our proposed algorithms <i>Simple</i> ap- proach and <i>GraphWalk</i> approach and the structural comparison algorithms	35
4.5	For the test set, the confusion matrix obtained by our <i>Simple</i> ap- proach	35

4.6	For the test set, the confusion matrix obtained by our <i>Graph Walk</i> approach	36
-----	--	----

Chapter 1

Introduction

Cancer is one of the most lethal diseases especially in developed and developing countries. Although many tests are present for cancer screening, the routine practice for cancer diagnosis and grading is the histopathological examination of a biopsy, which includes examining biopsy tissues under a microscope and diagnosing cancer based on abnormal tissue formations. Although this examination is the gold standard in the current practice of medicine, it is subject to observer variability since it mainly relies on the visual interpretation of pathologists that heavily depends on their experience and expertise. To alleviate the observer variability, it has been proposed to use computational methods that extract mathematical features to represent histopathological tissue images and use these features for their classification.

The previously proposed automated cancer diagnosis systems typically use one of the following two approaches to define their descriptors: textural and structural. In the literature, there exist several studies that use different texture descriptors for automated cancer diagnosis and grading. The most commonly used descriptors are those that are defined on intensity/color histograms, which quantify the first order statistics of image pixels [1, 2, 3, 4], and cooccurrence matrices, which quantify the second order statistics among pixels [4, 5, 6, 7, 8]. In addition to these, many studies make use of wavelets to define their features.

Examples include the descriptors defined on multiwavelet coefficients [9] and Gabor filter responses [10]. Fractal analysis is another method used for defining texture descriptors. In this analysis, fractal dimensions are frequently used as features [1, 11, 12, 13]. More recent studies use local binary patterns to define additional texture descriptors [14, 15, 16, 17, 18]. They are used to quantify a pixel according to spatial arrangement of its neighbors’ intensities with respect to its intensity. All these texture descriptors yield promising results. However, they are defined on pixels, directly using pixels’ intensity/color values. Thus, they are susceptible to pixel-level noise and variations that are typically observed in histopathological images.

To alleviate the negative aspects of pixel-level noise and variations, structural approaches have been proposed. These approaches are defined to model the spatial relationships of histological components to represent a tissue image. These approaches commonly construct a graph on these components and use graph descriptors for image classification. Earlier studies construct their graphs on only nucleus tissue components using different techniques such as Delaunay triangulations [7, 10, 19, 20, 21], minimum spanning trees [19, 22, 21], and probabilistic graph generations [23]. In the recent study of our research group [24], we construct a graph on tissue components of different types and color graph edges based on the types of their end nodes. Different than the descriptors proposed in this thesis, these previous structural approaches usually use a global graph representation for the entire image and extract global graph descriptors for its quantification.

1.1 Contribution

In this thesis, we propose a new algorithm for effective and robust representation and classification of images of histopathological colon tissues stained with hematoxylin-and-eosin. In the proposed algorithm, our main contributions are the introduction of a set of new texture descriptors, which we call local object patterns, to model composition of histological components in a tissue image and

the use of this descriptor set to define the visual words of the bag-of-words representation of the image. To this end, we decompose the image into component objects of multiple types and define texture of these objects using the idea of local binary patterns [25]. However, as opposed to local binary patterns defined at the pixel-level, we define local object patterns on the objects at the component-level. Particularly, local binary patterns are defined to quantify a pixel by constructing a binary string from the spatial arrangement of its neighbors' relative intensities. On the other hand, we define our local object patterns to quantify an object by specifying a set of neighborhoods with different locality ranges and constructing a string based on how the object's neighbors arrange in an order in each of these local neighborhoods. Our texture definition proposed in this thesis mainly differs from the previous texture-based tissue classification studies in the following aspect: It defines its texture descriptors on higher-level component objects instead of defining them at the pixel-level.

In the proposed algorithm, the defined local object patterns are used to characterize the tissue objects, the type distribution of which is used to construct bag-of-words representations of the tissue. In this thesis, we implement two algorithms for this construction. The first approach (*Simple approach*) uses the type distribution of the objects to construct a single bag-of-words representation. The second one (*GraphWalk approach*) uses the type distributions of the object subsets obtained through graph walking to construct multiple bag-of-words representations and combines them by voting. Different than the previous approaches, this thesis uses the object distributions whose types are assigned by making use of their local object patterns. Moreover, the use of graph walks to obtain multiple object subsets is another contribution of this thesis.

The algorithms proposed by this thesis are tested on 3236 microscopic colon tissue images. The experiments reveal that our proposed texture descriptors are effective to obtain better classification accuracies compared to previous texture definitions.

1.2 Outline of the Thesis

This thesis is structured as follows. In Chapter 2, we give details of the background information about the problem domain and summarize the previous computational methods used for automated cancer diagnosis and grading. In Chapter 3, we present our proposed local object pattern algorithm, providing the details of decomposing an image into components, defining local object patterns on these components, and using these patterns for image classification. In Chapter 4, we explain the dataset, and discuss the experimental results. Finally, in Chapter 5, we provide the summary of the thesis and discuss its future research directions.

Chapter 2

Background

In this chapter, we briefly give domain description of colon tissues and colon cancer. Then, we present the approaches that are used in the literature for automated cancer diagnosis.

2.1 Domain Description

This thesis focuses on representation of colon tissues and classification of colon adenocarcinoma. Colon tissues contain hierarchical structures called glands. These glands are formed of epithelial cells lined up around a lumen, which is an inner open space of a tubular structure and absorbs water and minerals while secreting mucus. Besides these gland structures, colon tissues contain stroma, which is a connective tissue and contains non-epithelial cells. These basic parts of a colon tissue are shown on an example image in Figure 2.1.

Colon cancer is one of the four major cancer sites [26], which results in uncontrolled cell growth in colon tissues. Colon adenocarcinoma is the cancer type that accounts for 90-95 percent of all colorectal cancers. Colon adenocarcinoma originates from epithelial cells forming the glandular structures in colon tissues. Thus, it causes changes in the glandular structures. Figure 2.2 shows normal

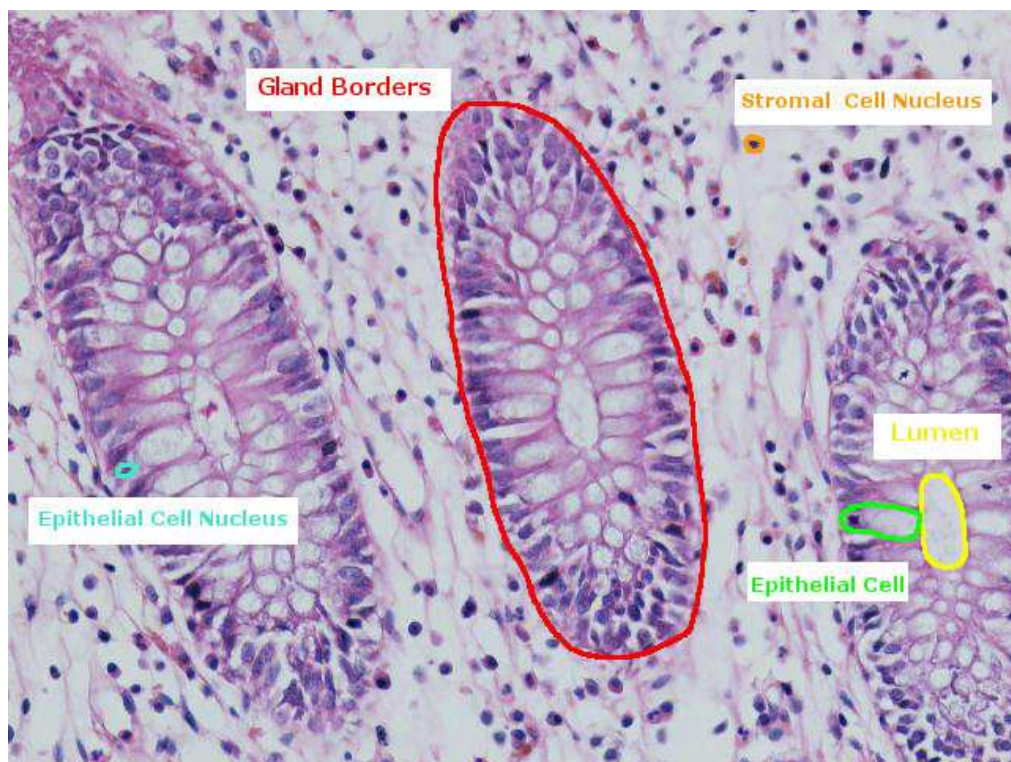


Figure 2.1: An image of a normal colon tissue stained with hematoxylin and eosin, which is the routinely used technique to stain biopsies in hospitals.

and adenocarcinomatous colon tissues. As shown in this figure, gland structures deform with the existence of cancer. This deformation is less and the glands are still differentiated in low-grade cancerous tissues whereas the deformation increases and the glands are only poorly differentiated in high-grade cancerous tissues. In this thesis, we focus on classifying tissue images into three classes: normal, low-grade cancerous, and high-grade cancerous.

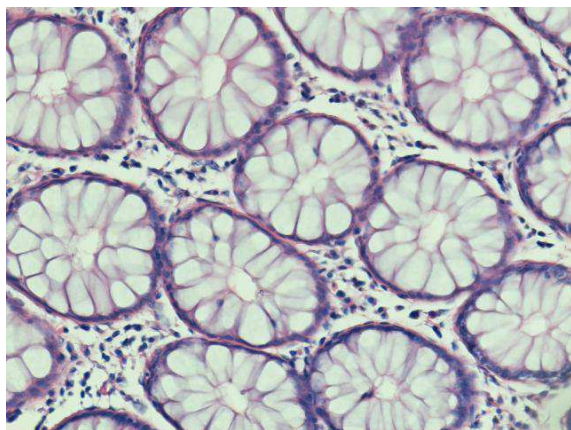
The final diagnosis of colon adenocarcinoma is done via histopathological examination of colon biopsies. In this process, a small amount of sample tissue (biopsy) is taken from the body with a special instrument and then is fixed, cut into thin pieces, and stained for microscopic examination. The staining technique that is mostly used in hospitals is called hematoxylin-and-eosin (H&E). In this staining, hematoxylin stains cell nuclei blue-purple and eosin stains proteins and other cellular elements in a tissue pink-red; background remains colorless [27]. The example tissue images obtained with this staining are shown in Figure 2.1 and Figure 2.2.

2.2 Automated Cancer Classification

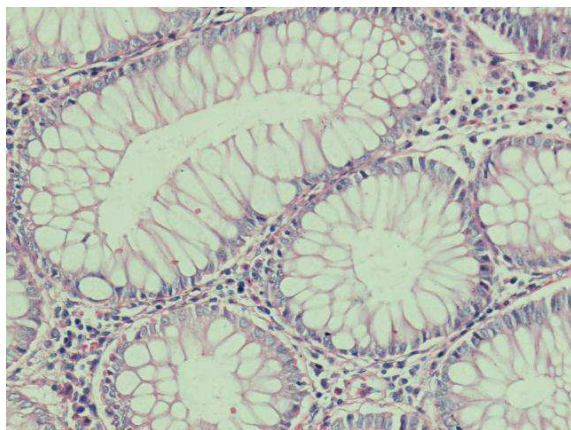
In this section, we will explain the previous methods that are used for automated diagnosis and grading of cancer. These methods can be mainly grouped into two: textural and structural. We will explain these methods in the following subsections.

2.2.1 Textural Methods

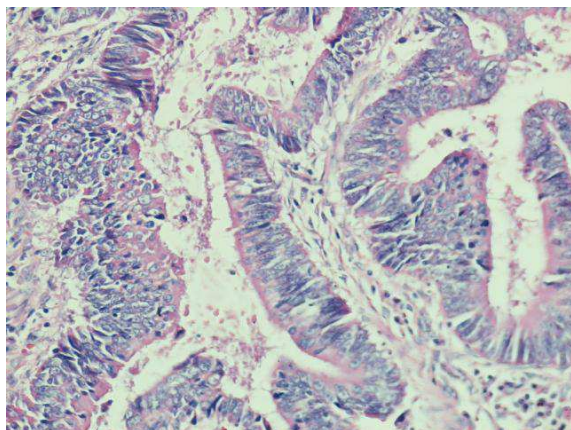
Textural methods are essential in the image analysis based on local spatial variations of intensity or color. These methods aim to draw conclusion for an unknown image by using a known texture. Textures are obtained from images using various texture feature extraction methods such as cooccurrence matrix features, fractal dimensions, run-length features, wavelet features, and entropy [28].



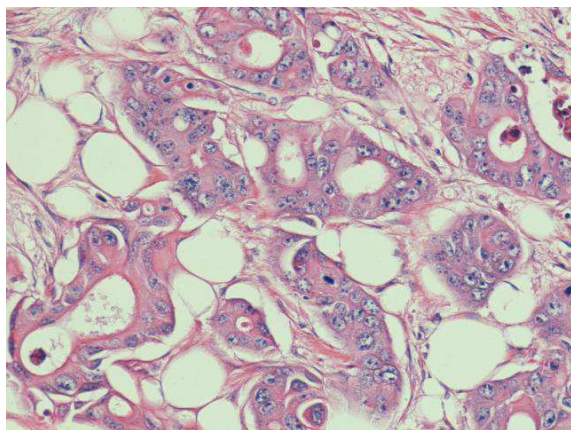
(a)



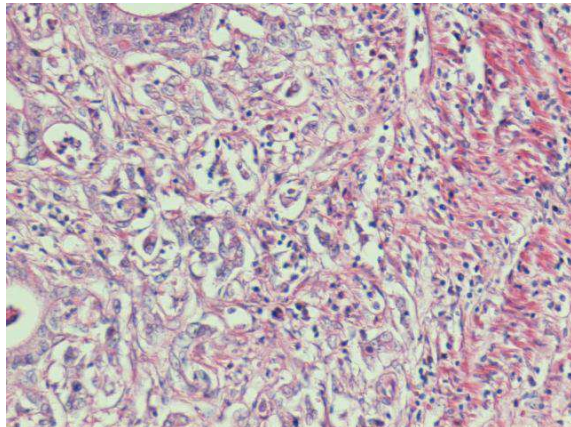
(b)



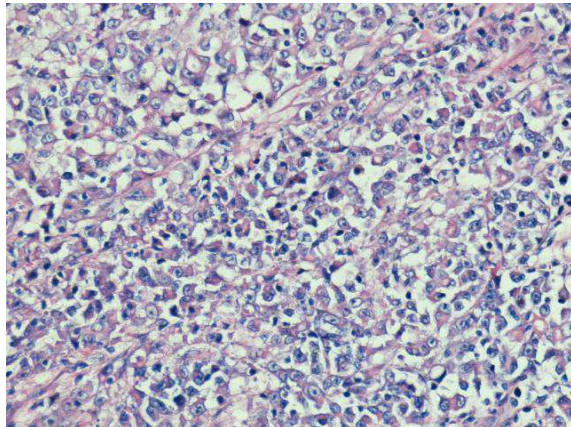
(c)



(d)



(e)



(f)

Figure 2.2: Examples of colon tissue images: (a)-(b) normal tissues, (c)-(d) low-grade cancerous tissues, and (e)-(f) high-grade cancerous tissues.

2.2.1.1 Color and Intensity Histogram Features

Histograms characterize an image according to its color distribution or intensity. The color histogram is used for cancer diagnosis [1, 2, 3]. It uses joint probabilities of intensities in the red, green, and blue color channels. It divides the image into bins and stores the pixels of each color channels. It can be formalized as

$$h_{R,G,B}(a, b, c) = N \cdot Prob(R = a, G = b, B = c) \quad (2.1)$$

where R, G, and B are the three color channels and N is the number of pixels in the image.

Intensity histogram illustrates the frequency of pixels in a grayscale image with L levels of intensity found in the image. It is a discrete function $h(r_k) = n_k$ where r_k is the k^{th} gray level and n_k is the frequency of pixels in the gray level r_k in range [0 - L]. The mean, standard deviation, skewness, kurtosis, and entropy (Table 2.1) are extracted from the histogram probability density function [4]. $p(r_k) = \frac{h(r_k)}{M}$ where M is the number of pixels, and $p(r_k)$ should satisfy the following properties.

$$\left. \begin{array}{l} p(r_k) \geq 0 \\ \sum_{g=0}^L p(g) = 1 \end{array} \right\}$$

Table 2.1: The most commonly used intensity histogram features

Mean	$\sum_k p(r_k) \cdot r_k$
Standard Deviation	$\sum_k (r_k - mean)^2 \cdot p(r_k)$
Skewness	$\sum_k (r_k - mean)^3 \cdot p(r_k)$
Kurtois	$\sum_k (r_k - mean)^4 \cdot p(r_k)$
Entropy	$-\sum_k p(r_k) \cdot \log_2 p(r_k)$

2.2.1.2 Cooccurrence Matrix Features

The most widely used features for textural analysis are Haralick features extracted from cooccurrence matrices [29]. They are calculated to quantify the relationship among two pixels that cooccur at a specific distance and direction. For a given direction Q and distance d , cooccurrence matrix keeps the frequency of occurrences of gray levels i and j . The features extracted from a cooccurrence matrix are summarized in Table 2.2 where $p(i,j)$ is the frequency of cooccurrence of gray levels i and j , μ is the mean, and σ is the standard deviation. Although, 14 texture features are initially proposed, only four of them, which are angular second moment, contrast, correlation, and entropy are widely used. A cooccurrence matrix is a second order statistics that defines how often different variation of pixel gray levels cooccur in an image. Cooccurrence matrix denotes various characteristic of the spatial distribution of the gray levels in entire images so the features are used in tissue analysis [4, 6, 7, 8].

2.2.1.3 Run-Length Matrix Features

In cancer diagnosis and grading, higher order statistics [30] extracted from run-length matrices are also used [31, 32, 33]. Run-length is modeled as a pattern of the length of the scanned line of each image in pixels. A primitive is defined as a continuous set of maximum number of pixels both in the same gray levels and in the same direction [34]. A number of primitives of all directions $p(a,r)$, (where length r with a gray level in an $M \times N$ image) is used for extracting the features. Galloway [30] extracts five features whereas Chu [35] defines two more new features. Finally, Dasarathy and Holder [36] extend the features by defining new four features. Table 2.3 illustrates the mostly used run-length matrix features for cancer diagnosis and grading. In this table, $K = \sum_{a=1}^L \sum_{r=1}^{N_r}$ is the total number of runs, L is the number of gray levels, and N_r is the the maximum length.

Table 2.2: Cooccurrence matrix features

Angular Second Moment	$\sum_i \sum_j p(i, j)^2$
Contrast	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\}, i - j = n$
Correlation	$\frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Sum of Squares : Variance	$\sum_i \sum_j (i - \mu)^2 p(i, j)$
Inverse Difference Moment	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$
Sum Average	$\sum_{i=2}^{2N_g} i p_{x+y}(i)$
Sum Variance	$\sum_{i=2}^{2N_g} (i - f_8)^2 p_{(x+y)}(i)$
Sum Entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log \{p_{x+y}(i)\} = f_8$
Entropy	$-\sum_i \sum_j p(i, j) \log(p(i, j))$
Difference Variance	$\sum_{i=0}^{N_g-1} i^2 p_{x-y}(i)$
Difference Entropy	$\sum_{i=0}^{N_g-1} p_{x-y}(i) \log \{p_{x-y}(i)\}$
Info. Measure of Correlation 1	$\frac{HXY - HXY_l}{\max \{HX, HY\}}$
Info. Measure of Correlation 2	$(1 - \exp[-2(HXY_2 - HXY)])^{\frac{1}{2}}$

Table 2.3: The mostly used run-length matrix features

Short primitive emphasis	$\frac{1}{K} \sum_{a=1}^L \sum_{r=1}^{N_r} \frac{p(a, r)}{r^2}$
Long primitive emphasis	$\frac{1}{K} \sum_{a=1}^L \sum_{r=1}^{N_r} p(a, r) r^2$
Gray level uniformity	$\frac{1}{K} \sum_{a=1}^L [\sum_{r=1}^{N_r} p(a, r) r^2]^2$
Primitive length uniformity	$\frac{1}{K} \sum_{a=1}^L [\sum_{r=1}^{N_r} p(a, r)]^2$
Primitive percentage	$\frac{K}{\sum_{a=1}^L \sum_{r=1}^{N_r} p(a, r)} = \frac{K}{MN}$

2.2.1.4 Law's Texture Energy Measures

It has been proposed to use Laws' texture energy measures [37] to describe textures for cancer diagnosis and grading [38]. The Laws' texture features are similar to the Haralick's [29] co-occurrence matrix features [29]. They calculate texture energies in the spatial domain. Given one-dimensional kernels merged into convolution masks, which output the energy image and every pixel located at the center of the local window's $l(i, j)$ are replaced with the absolute value in the filter window $f(i, j)$ where n is the size of the mask as given below:

$$s(i, j) = \frac{1}{(2xn + 1)^2} \sum_{k=i-n}^{i+n} \sum_{l=j-n}^{j+n} |f(k, l) - l(i, j)| \quad (2.2)$$

2.2.1.5 Fractal Analysis

Fractal geometry is used for cancer diagnosis and grading [1, 12, 13]. The fractal dimension is the most commonly used feature in fractal analysis and computes the self-similarity property. Although there exist various fractal dimension algorithms, the box-counting method [39] is commonly used due to its cost efficiency, high accuracy, and implementation ease [13]. The formula for the box-counting method is given in Equation 2.3.

$$D_B = -\lim_{\epsilon \rightarrow 0} \frac{\log N_\epsilon(S)}{\log(\epsilon)} \quad (2.3)$$

2.2.1.6 Multiwavelet Features

Wavelets create whole representation of signals by using all of the sub-band decomposition and allow the image decomposition into various frequency sub-bands. Wavelets utilize one scaling function, whereas multiwavelets use more than one scaling function. In cancer diagnosis and grading, wavelets are employed to extract features from the wavelet coefficients (like entropy and energy) [9] and from Gabor filter responses [7].

2.2.1.7 Local Binary Patterns

Local binary patterns define the relationship between a pixel and its neighborhood pixels. Initially, Ojala et al. [40] proposed a method where neighbor pixels whose intensities are higher than or equal to the value of the center pixel are labeled with 1 and the others are labeled with 0. The method is improved by considering the neighbors which are radius r distance far away from the center pixel [25]. The neighbors are counted clockwise to get the binary value as follows:

$$LBP_{p,r} = \sum_{n=0}^{p-1} s(x_{r,n} - x_{0,0})2^n, \quad \left. \begin{array}{l} 1, x \geq 0 \\ 0, x < 0 \end{array} \right\} = s(x)$$

where $x_{0,0}$ is center pixel and p is the number of neighbor pixels around radius circle r .

In cancer diagnosis, local binary pattern texture features are usually combined with other features by various methods [14, 15, 16, 17, 18].

2.2.2 Structural Methods

Textural approaches may suffer from noise and variation commonly observed at the pixel-level of tissue images. On the other hand, structural approach define their features at the component level to mitigate the pixel-level problems. In automated cancer diagnosis, graph based methods are widely used as structural approaches. In these methods, nodes are mostly the nuclear components and edges are defined for encoding the spatial information among the nodes. Voronoi diagrams, Delaunay triangulations, minimum spanning trees, probabilistic graphs, weighted graphs, and color graphs are the most common graph generation algorithms. These generation methods will be discussed in the following subsections.

2.2.2.1 Voronoi Diagrams

The Voronoi diagram of a point set S divides the n points in the set S into n regions where $p \in S$ contains all points in the plane such that p is the nearest site [41]. Figure 2.3 illustrates the Voronoi diagram of randomly selected 20 points. The features extracted from a Voronoi diagram include the area, roundness, aspect ratio, circularity, and the number of sides. They are commonly used with the features extracted from Delaunay triangulation [7, 19, 20, 21].

2.2.2.2 Delaunay Triangulations

A triangulation of S is a planar graph with a vertex set where all the bounded faces are triangles and S is a set of n points. The Delaunay triangulation of S is a dual graph of the Voronoi diagram where the edges are straight lines and each vertex is located in the set S . Delaunay triangulation are used in many cancer diagnosis and grading algorithms [19, 42, 43]. Triangulations are characterized by using structural features such as the area, edge length, degree of nodes, distance to the nearest neighbor, eccentricity, clustering coefficient, shortest paths between nodes, and diameter. Figure 2.4 shows the Delaunay triangulation of the same randomly selected 20 points used in the construction of the Voronoi graph given

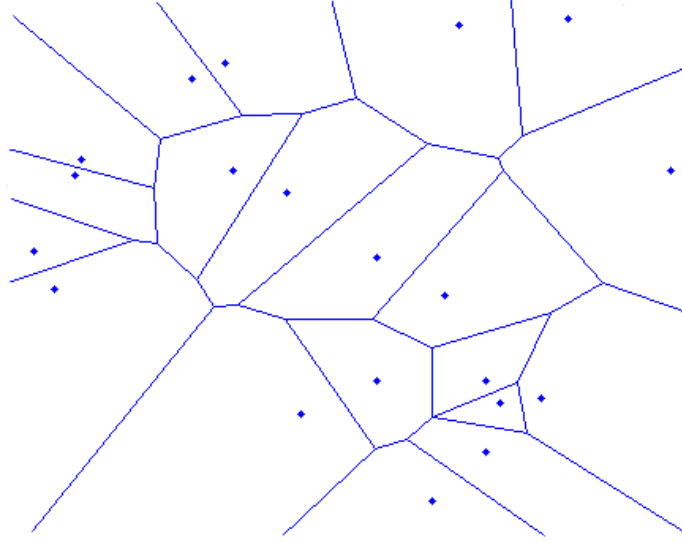


Figure 2.3: The Voronoi diagram for randomly selected 20 points

in Figure 2.3. Note that it is also possible to use Gabriel's graph to extract such features [44]. Such as a Gabriel graph is shown in Figure 2.5.

2.2.2.3 Minimum Spanning Trees

In graph theory, a spanning tree T , is a sub-graph of a connected, undirected graph $G = (W, E)$ that includes every vertex of the graph G . A minimum spanning tree of graph G is the spanning tree, for which the sum of edge weights is minimized. Like Voronoi diagrams and Delaunay triangulations, structural features are extracted from minimum spanning trees. These are the edge length, degree of nodes, distance to the nearest neighbor, fractal dimension, eccentricity, clustering coefficient, and diameter [19, 21, 22]. Figure 2.6 demonstrates the minimum spanning tree of the same randomly selected 20 points.

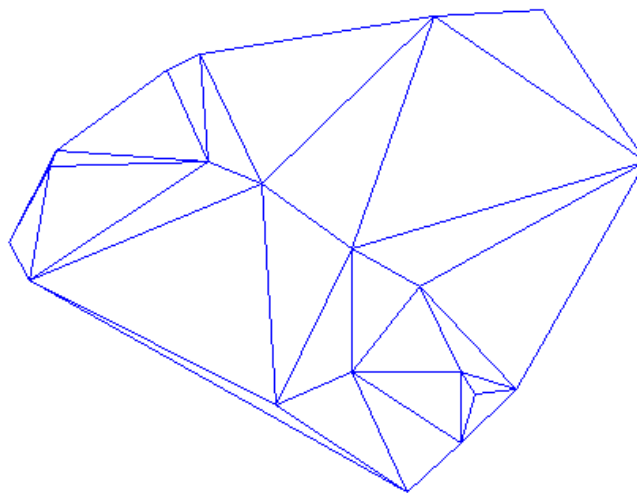


Figure 2.4: The Delaunay triangulation for the 20 random selected points

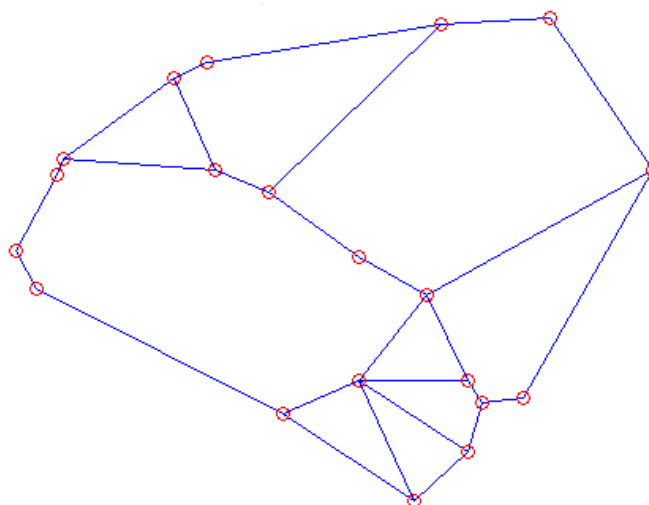


Figure 2.5: The Gabriel's graph for the same randomly selected 20 points

Chapter 3

Methodology

Our proposed algorithm introduces a new texture descriptor, which we call local object patterns, to model tissue images and uses these descriptors for tissue image classification. To this end, it decomposes a tissue image into its histological components, characterizes them with the newly introduced local object pattern descriptors, and uses this characterization for classification of tissue images. In the following sections, the details of the proposed algorithm are provided.

3.1 Tissue Image Decomposition

We model a tissue image I by approximately representing its histological components with a set of circular objects $O(I) = \{o_i\}$. We represent each object o_i by its coordinates (x_i, y_i) and its type $t_i \in \{purple, pink, white\}$. These types correspond to the three main colors in a hematoxylin-and-eosin stained tissue. Particularly, cell nuclei correspond to purple; stroma, stromal cells' cytoplasms, and mucin-poor epithelial cells' cytoplasms correspond to pink; and lumina and mucin-rich epithelial cells' cytoplasms correspond to white. Since there are multiple components corresponding to the same type, we hereinafter refer them to as purple, pink, and white, to keep the thesis simpler and easier to read.

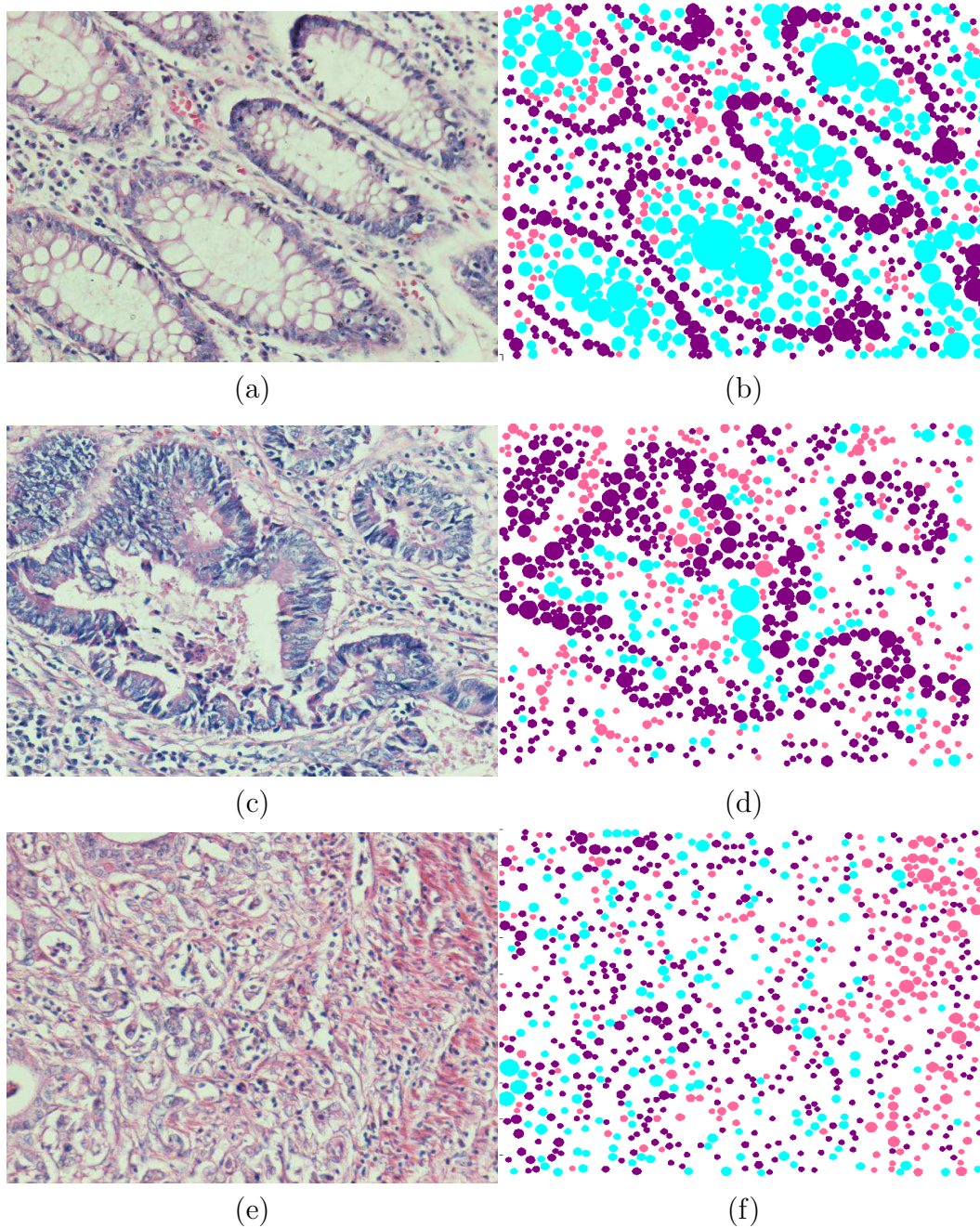


Figure 3.1: Examples of tissue images and their located objects. In these images, (a) a normal tissue image, (c) a low-grade cancerous tissue image, (e) a high-grade cancerous tissue image and (b)(d)(f) the objects located on these tissue images. Here purple, pink, and white objects are represented with purple, pink, and cyan, respectively.

In order to define the circular object set $O(I) = \{o_i\}$, we first separate hematoxylin and eosin channels of the image I by applying a color deconvolution method [45]. Then, we quantize pixels into three groups (pink, white, and purple which are the main colors in the hematoxylin and eosin stained tissues) according to the color convolution. Particularly, we assign pixel p_i into one of these groups as follows:

$$p_i = \begin{cases} \text{purple} & \text{if } h_i \leq h_{avg} \\ \text{pink} & \text{else if } e_i \leq e_{avg} \\ \text{white} & \text{otherwise} \end{cases}$$

where h_i and e_i are the hematoxylin and eosin component values of the pixel p_i and h_{avg} and e_{avg} are the average values of the hematoxylin-and-eosin component values over all pixels. After pixels are labeled, we apply the circle fit algorithm [46] to each group's pixels to locate the objects. In this algorithm, the objects are located if radii are greater than the threshold radius r_{min} . Figure 3.1 illustrates the objects located on example tissue images.

In our model, we use an approximate representation instead of finding exact locations of histological components because their exact localization gives rise to a quite difficult segmentation problem. Thus, there may be one-to-one or many-to-one relation between objects and components. For example, a purple object usually corresponds to a single nucleus, whereas a group of white objects that form a clique corresponds to a lumen region. The proposed local binary patterns are also effective to model such many-to-one relations.

3.2 Local Object Patterns

For object o_i , the n^{th} local object pattern $n\text{-LOP}(o_i)$ is defined as follows: we first find the distance from the coordinate (x_i, y_i) of o_i to the coordinate of every other object in the object set $O = \{o_1, \dots, o_i, \dots, o_k\}$ and select the n nearest neighbors of o_i to get the ordered neighbor set $N(o_i) = \{o_{i1}, \dots, o_{ij}, \dots, o_{in}\}$. After that, we form a binary string $B(o_i) = \{b_{i1}, \dots, b_{ij}, \dots, b_{in}\}$. In this string;

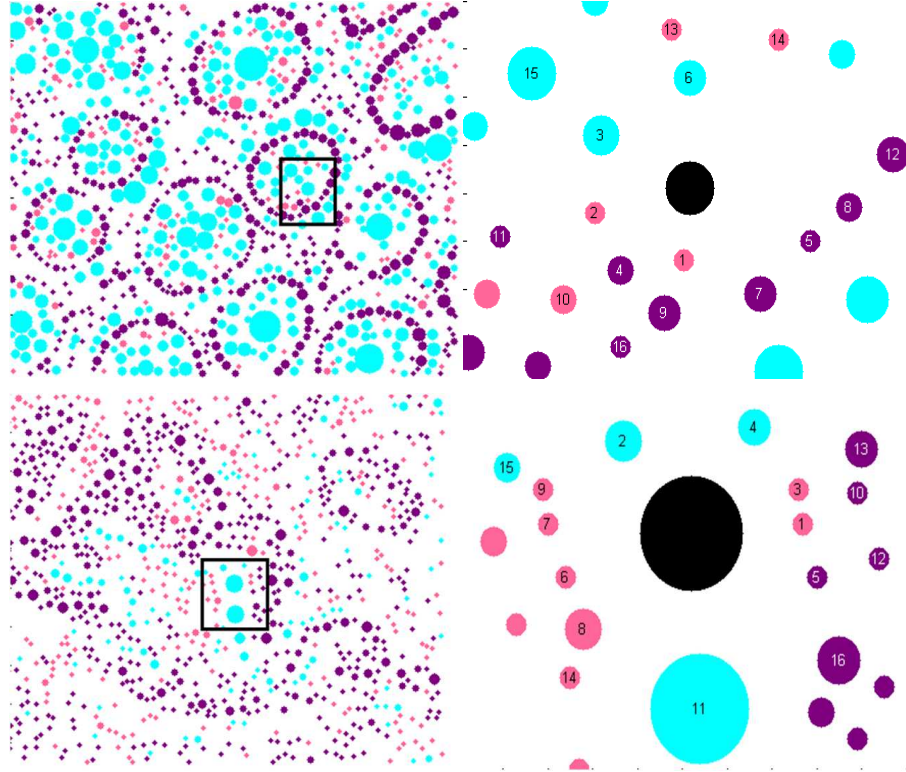
$$b_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in \{purple\} \\ 0 & \text{if } t_{ij} \in \{pink, white\} \end{cases}$$

where t_{ij} is the type of the selected neighbor o_{ij} . Then, we define 2^j -LOP(o_i) as the decimal equivalent of the binary string $B(o_i)$. Note that this descriptor provides rotation invariance since objects are ordered based on their distances to object o_i and its value does not change with arbitrary rotations of the image.

We define local object patterns for an object to quantify the spatial arrangement of its neighbors' types found in a local neighborhood. In our model, we extract a set of $(m+1)$ patterns using different neighborhoods. Particularly, this set includes $S = \bigcup_{j=0}^m 2^j$ -LOP.

Figure 3.2 illustrates extraction of local object patterns for the objects shown in black. We select these objects such that they both belong to luminal regions; we crop these regions from normal and cancerous tissue images as shown in Figure 3.2. This illustration shows that although lower-order patterns are the same for the two selected objects, their higher-order patterns show differences, which can be used to differentiate these objects.

By using the local object pattern descriptors, we define the new object types as follows: For each original type $t_i \in \{purple, pink, white\}$, we separately cluster objects of the corresponding type into k groups running the k-means algorithm on local object patterns of these objects. Thus, we learn k clustering vectors $V_{purple} = \{v_1, v_2, \dots, v_k\}$ for the purple type, k clustering vectors $U_{pink} = \{u_1, u_2, \dots, u_k\}$ for the pink type, and k clustering vectors $W_{white} = \{w_1, w_2, \dots, w_k\}$ for the white type. Then, for a given image, we relabel each object o_i with a new type t'_i based on its original type t_i and the corresponding set of the clustering vectors V_{purple} , U_{pink} or W_{white} . Since components (objects) of normal and cancerous tissue images show different neighbor distributions, they are expected to be relabeled with different types t'_i . Thus, we use the distribution of these new types to represent an image. In our work, we use the distribution of the new types in two different ways. We will explain these ways



	Normal Tissue	Cancerous Tissue
1-LOP	$\text{decimal}(\langle 0 \rangle)$	$\text{decimal}(\langle 0 \rangle)$
2-LOP	$\text{decimal}(\langle 00 \rangle)$	$\text{decimal}(\langle 00 \rangle)$
4-LOP	$\text{decimal}(\langle 1000 \rangle)$	$\text{decimal}(\langle 0000 \rangle)$
8-LOP	$\text{decimal}(\langle 1101\ 1000 \rangle)$	$\text{decimal}(\langle 0001\ 0000 \rangle)$
16-LOP	$\text{decimal}(\langle 1000\ 1101\ 1101\ 1000 \rangle)$	$\text{decimal}(\langle 1001\ 1010\ 0001\ 0000 \rangle)$

Figure 3.2: Extracting local object patterns for the objects shown in black. Here m is selected as 4, thus $S = \bigcup_{j=0}^m 2^j\text{-LOP}$. Sixteen-nearest neighbors of the selected objects are indicated on the examples with their orders.

in the following section.

3.3 Tissue Classification

After characterizing the objects of an image I , we use this characterization to represent the image. For that, we use the bag-of-words representation on the frequency of objects' new types. Here we use two approaches to obtain the bag-of-words representation.

In the *Simple Approach*, we use the new types of all objects in the image to extract the bag-of-words representation and classify the image using a linear kernel support vector machine (SVM) classifier. In our work, we use the SVM library given in [47]. Here since our problem involves more than two classes, this implementation uses one-against-one strategy for this multiclass classification. This strategy constructs $k(k - 1)/2$ classifiers, each of which differentiates one class from another. Then the decisions of these classifiers are voted to obtain the final class.

In the *GraphWalk Approach*, we obtain multiple subsets of the objects and separately use the frequency of their types to obtain multiple bag-of-words representations of the same image. Here we use this approach to explore the following: when all objects are used to create a bag-of-words representation, this gives a global feature set of the entire image. However, in tissue images, there may exist some local subregions that may be more important than the others. By selecting object subsets that correspond to such local regions, we aim to extract feature sets corresponding to these potentially important subregions.

In this second approach, we first construct a graph on the entire objects by using Delaunay triangulation and then employ the breadth first search (BFS) algorithm to select an object subset. We finally use the frequency of the new types of these selected objects to construct a bag-of-words representation. Particularly, the BFS algorithm traverses graph nodes, which correspond to the objects, level-by-level starting from an initial node (object). In our algorithm, we start the

BFS algorithm from each of the N largest white objects and obtain an object subset. Here we terminate the BFS algorithm, after it traverses L objects in the graph. Here it is worth noting that the number L of visited objects is less than the number of the entire objects in the image. Thus, the selected object subsets are expected to cover a smaller local region in the tissue image. In our work, we classify each of these bag-of-words representations with linear kernel SVMs and combine the decisions of these SVMs through simple voting.

Chapter 4

Experimental Results

In this chapter, we first give the details of our dataset. Then, we explain the methods that we use in our comparisons. Later, we explicate the parameters of our algorithms. After that, we give our results and compare them with the comparison methods to understand how efficient and accurate our proposed algorithms are. Finally, we provide the effects of the selection of the model parameters to classification accuracy.

4.1 Dataset

Our dataset contains 3236 microscopic images of colon tissues stained with hematoxylin-and-eosin. The images are taken from the 258 randomly selected patients from the Pathology Department in Hacettepe University School of Medicine. They are acquired by a Nikon Coolscope Digital Microscope using 20 $\times$ microscope objective lens and 480×640 image resolution.

Images are randomly divided into two groups as training and test sets. The 1644 images of randomly selected patients constitute the training set and 1592 images of the rest of the patients constitute the test set. Each image in these sets is labeled as normal, low-grade cancerous, or high-grade cancerous. The

training set contains 510 normal, 859 low-grade cancerous, and 275 high-grade cancerous tissues of 129 patients. The test set contains 491 normal, 844 low-grade cancerous, and 257 high-grade cancerous tissues of the remaining 129 patients.

4.2 Comparisons

We use two groups of approaches in our comparison: textural and structural. These comparison approaches are given in the following subsections.

4.2.1 Textural Approaches

First, we compare our algorithm that defines texture descriptors at the component level with the textural approaches that define their descriptors at the pixel level. These are intensity histograms, gray-level cooccurrence matrices, and local binary patterns. These approaches use linear kernel SVMs in their classification.

4.2.1.1 Intensity Histogram Approach

First order histogram features are derived from the gray-level intensities of image pixels. They include mean, standard deviation, kurtosis, and skewness. To reduce noise and small intensity variations, pixel intensities are grouped into N bins [4].

We also implement the grid based variant of the *Intensity Histogram* approach. Here we decide to use a grid-based variant since it is commonly difficult to find a constant texture over an entire image as the tissue image may contain sub-regions irrelevant to classification. In this approach, we divide the image into fix-size grids, extract a histogram on each grid, calculate descriptors on the grid histograms, and average these descriptors all over the image.

4.2.1.2 Cooccurrence Matrix Approach

Second order statistics are calculated on the gray level intensities of image pixels. In our comparisons, we use textural features that are extracted from cooccurrence matrices at eight orientations. These features include the angular moment to model homogeneity of a tissue image, the entropy to model randomness, the contrast and difference moment to model local variations, the correlation function to model linearity of gray-level dependencies, the inverse difference moment to model local homogeneity, the dissimilarity to measure the dissimilarity between pixels [5].

Likewise, we also use a grid-based variant of the *Cooccurrence Matrix* approach. Similarly, we divide the image into fixed size grids, extract cooccurrence matrix features from each of these grids and average them all over the grids [48].

4.2.1.3 Local Binary Pattern Approach

The *Local Binary Pattern* descriptors include histogram frequencies. We compute this histogram on the outputs of a uniform local binary pattern (LBP) operator applied on image pixels. For each pixel, the *LBP* operator [25] outputs a binary string by comparing the pixel’s gray-scale intensity with those of its eight neighbors; it outputs 1 if its intensity is lower and 0 otherwise. It then assigns the pixel to a histogram bin based on the number of consecutive 1’s in this binary string. This operator is called uniform if it constructs the histogram on only the pixels whose binary strings contain at most two bitwise 0/1 transitions in their circular chain. We calculate an additional bin for keeping frequencies of pixels with non-uniform strings. In our experiments, we extract these descriptors from the histogram constructed on all pixels. Here we did not implement its grid based variant because calculating histograms on pixels of equal-sized grids and averaging their histogram frequencies is equivalent to calculating a histogram on all pixels and using its frequencies.

4.2.1.4 Pixel Based Approach

This is the pixel-based counterpart of our algorithm. This approach follows exactly the same steps of our algorithm except its descriptor definition step. Particularly, it decomposes a tissue image into a set of circular objects, defines descriptors on the objects, clusters the objects based on their descriptors to find their new types, and uses the new types' frequency in a linear kernel SVM classifier. Here different than our proposed algorithm, which uses local object patterns as the descriptors, the *Pixel Based* approach uses local binary patterns. To this end, it locates a square window at the center of each object and calculates local binary patterns of this window to find the descriptors of the object. We use this comparison in our experiments to understand the effectiveness of defining component-level local object patterns.

4.2.1.5 Resampling-based Markovian Model

Additionally, we use the *Resampling-based Markovian Model* (RMM) that we previously implemented in our research group [48]. The *RMM* obtains multiple samples of an image, labels each sample using discrete Markov models, and votes the samples' labels to classify the image. To obtain an image sample, it generates a sequence on the randomly selected points, which are characterized by texture descriptors and ordered based on proximity. These descriptors include the histogram of quantized pixels and the J-value texture measure.

4.2.2 Structural Methods

Next, we compare our algorithm with previous structural approaches. These approaches construct graphs on tissue components and extract features on these constructed graphs. Likewise, these approaches use linear kernel SVM classifiers. The details of these methods are given below:

4.2.2.1 Delaunay Triangulation

This approach defines its graph constructing Delaunay triangulation on nuclear (purple) objects located by the circle fit algorithm. Then it extracts global features from this Delaunay triangulation. These features include the average, standard deviation, minimum-to-maximum ratio, and disorder of edge lengths and triangle areas, as well as the average degree, average clustering coefficient, and diameter of the entire Delaunay graph [10].

4.2.2.2 Color Graph

This is similar to the *Delaunay Triangulation* approach except that it constructs its graph on all types of tissue components. Particularly, it constructs Delaunay triangulation on all objects but colors the triangle edges based on the types of the end nodes. Then, it extracts colored version of the global features including the average degree, average clustering coefficient, and diameter [24].

4.2.2.3 Hybrid Model

The last method is the *Hybrid Model* that we recently developed in our research group [49]. This model first represents an image with an attributed graph and defines smaller query graphs as a reference to normal gland structures. It then selects regions of the image whose subgraphs are most structurally similar to the query graphs based on graph edit distances. Using the graph edit distances of the selected regions as well as their texture descriptors, it classifies the image by a linear kernel SVM.

4.3 Parameter Selection

We used two algorithms to define bag-of-words representations: *Simple* approach, which employs all objects, and *GraphWalk* approach, which uses object sets obtained by graph walking. The parameters of these algorithms are explained in the following subsections. We also explain the parameters of the comparison algorithms.

In all algorithms, we use three fold cross validation on the training set for parameter selection. In order to do that, we list all possible values of each parameter and then test all possible parameter combinations in the list, and select the one that yields the highest cross-validation accuracy. The three fold cross validation method divides the training set into three equal parts. It trains the set with the two subsets and tests the classifier with the third one. It repeats this for three times in each of which the classifier is tested with a different subset. Note that we consider the highest average accuracy obtained on the test subsets.

4.3.1 Simple Approach

The *Simple* approach has three model parameters:

Minimum Radius r_{min}

The minimum circle radius of the objects in tissue images

Highest Degree m

The highest degree in determining how many nearest neighbors should be selected

Cluster Number k

The cluster number for grouping the objects of each original type in tissue images

In addition to these parameters, we have an additional parameter C for the SVM classifier with a linear kernel [47]. In our experiments, we used all possible

Table 4.1: For textural approaches, considered parameter values are listed. The parameter values selected by three fold cross-validation are indicated as bold.

Intensity Histogram	$C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \mathbf{200}, \dots, 1000\}$ Bin numbers $\in \{4, 8, \mathbf{16}, 32\}$
Intensity Histogram Grid	$C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, \mathbf{550}, \dots, 1000\}$ Bin numbers $\in \{4, \mathbf{8}, 16, 32\}$ Grid Size $\in \{\mathbf{10}, 20, 40, 80\}$
Cooccurrence Matrix	$C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, \mathbf{900}, 1000\}$ Bin numbers $\in \{4, \mathbf{8}, 16, 32\}$ Distance $\in \{\mathbf{5}, 10, 20, 40\}$
Cooccurrence Matrix Grid	$C \in \{1, 2, \dots, 9, 10, 20, \mathbf{30}, \dots, 90, 100, 150, \dots, 1000\}$ Bin numbers $\in \{4, \mathbf{8}, 16, 32\}$ Distance $\in \{5, \mathbf{10}, 20, 40\}$ Grid Size $\in \{10, 20, \mathbf{40}, 80\}$
Local Binary Pattern	$C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, 200, \dots, \mathbf{700}, \dots, 1000\}$
RMM	$C \in \{1, 2, 3, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, 1000\}$ winSize $\in \{10, 20, \mathbf{40}, 80\}$ StateNo $\in \{4, 8, 16, 32, \mathbf{64}\}$ SeqLen $\in \{10, 25, 50, \mathbf{100}, 150\}$ SeqNo $\in \{10, 25, 50, \mathbf{100}, 150\}$
Pixel Based Approach	$C \in \{1, 2, \dots, 9, 10, 20, \dots, \mathbf{60}, \dots, 100, 150, 200, \dots, 1000\}$ $r_{min} \in \{3, \mathbf{4}, 5\}$ $k \in \{5, 10, 20, \mathbf{30}\}$ winSize $\in \{10, 20, \mathbf{40}, 80\}$

Table 4.2: For structural approaches, considered parameter values are listed. The parameter values selected by three fold cross-validation are indicated as bold.

Delaunay Triangulation	$C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, \mathbf{900}, 1000\}$ Structing element size $\in \{\mathbf{3}, 5, 7, 9\}$ Circle area threshold $\in \{5, \mathbf{10}, \dots, 50\}$
Color Graph	$C \in \{1, 2, \mathbf{3}, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, 1000\}$ Structing element size $\in \{\mathbf{3}, 5, 7, 9\}$ Circle area threshold $\in \{5, \mathbf{10}, \dots, 50\}$
Hybrid Model	$C \in \{1, 2, 3, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, 1000\}$ $W \in \{10, 20, 40, \mathbf{60}, 80\}$ $N \in \{1, 5, \mathbf{10}, 20, 30, 40\}$

combinations of the following parameter sets $r_{min} \in \{3, 4, 5\}$, $m \in \{2, 3, 4, 5\}$, $k \in \{5, 10, 20, 30\}$ and $C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, 950, 1000\}$. In our experiments, we get the highest cross-validation accuracy for $r_{min} = 4$, $m = 4$, $k = 20$, and $C = 20$.

4.3.2 GraphWalk Approach

The *GraphWalk* approach has the following parameters:

Subset Number N

The number of object subsets selected by the breadth first search algorithm

Visited Object Number L

The number of visited objects during the breadth first search algorithm

Highest Degree m

The highest degree in determining how many nearest neighbors should be selected

Cluster Number k

The cluster number for grouping the objects of each original type in tissue images

This approach has also the parameter C for the SVM classifier [47]. In its parameter selection, we use the following combinations of the parameter sets $m \in \{2, 3, 4, 5\}$, $k \in \{60, 70, 120, 150\}$, $N \in \{25, 50, 75, 100\}$, $L \in \{25, 50, 75, 100\}$, and $C \in \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, 400, \dots, 1000\}$. We obtain the highest cross validation accuracy for $m = 4$, $k = 70$, $N = 25$, $L = 100$ and $C = 400$.

4.3.3 Comparison Algorithms

We also use three fold cross-validation on the training set to select the parameters of the comparison algorithms. The lists of the considered parameter values for

textural and structural approaches are given in Table 4.1 and Table 4.2, respectively. The selected parameters of each approach are shown as bold.

4.4 Results

We implement local object patterns to represent and classify tissue images. We characterize the objects using these local object patterns and construct bag-of-words representation using two approaches: *Simple* approach and *GraphWalk* approach. In order to understand the efficiency of these algorithms, we compare them with the previous textural and structural algorithms. The test results are given in the Table 4.3 and Table 4.4 for textural and for structural algorithms, respectively. These tables show that our proposed algorithms give high accuracies for all classes.

Our algorithm has a similar methodology to the *Pixel Based* approach and differentiate only in the definition of its descriptors. The *Pixel Based* approach uses pixel-based local binary patterns as descriptors whereas our methods use object-based local patterns. When we compare these algorithms, our proposed approaches give higher accuracies in all classes and surpass the *Pixel Based* approach in grading high-grade cancerous tissues. The results show that defining local object patterns instead of local binary patterns increases the efficiency of tissue image classification.

The *Intensity Histogram*, *Cooccurrence Matrix*, and *Local Binary Pattern* approaches extract global texture descriptors on the entire tissue images and their grid-based variants make use of grids to extract their descriptors. The *RMM* approach extracts its pixel-based texture descriptors locally defined for the selected points. On the other hand, the proposed algorithm extracts texture descriptors for each object in a local neighborhood defined by the distance from this object to its 2^m -nearest neighbor. The results show that using object-based textures is more effective to obtain higher accuracies. The grid based variants improve results; however, they are still lower than the results of the proposed algorithm.

The *Delaunay Triangulation* and *Color Graph* approaches also use the circular objects and they construct graphs on these objects and extract global properties from these graphs. The *Hybrid Model* approach uses local graphs for the selected regions. Our *Simple* approach gives higher accuracies compared to these structural algorithms whereas our *GraphWalk* approach gives higher results than the *Delaunay Triangulation* and *Color Graph* approaches and similar results with the *Hybrid Model* approach. However, when we look at the accuracies of separate classes, we observe that the *GraphWalk* approach gives more balance results for the discrimination of different classes.

Table 4.3: Test set results obtained by our proposed algorithms *Simple* approach and *GraphWalk* Approach and the textural comparison algorithms

	Normal	Low-Grade	High-Grade	Overall
Simple Approach (LOPs)	95.32	92.54	90.27	93.03
GraphWalk Approach (nLOPs)	93.68	91.23	90.66	91.90
Intensity Histogram	80.65	69.55	70.04	73.05
Cooccurrence Matrix	83.10	81.64	77.82	81.47
Local Binary Pattern	92.67	73.46	80.54	80.53
Intensity Histogram Grid	78.82	74.17	78.60	76.32
Cooccurrence Matrix Grid	87.58	84.12	85.60	85.43
RMM	95.64	87.77	88.56	90.32
Pixel Based Approach	94.50	90.17	76.65	89.32

We propose two algorithms for defining bag-of-words representations: *Simple* approach and *GraphWalk* approach. The *Simple* approach defines a bag-of-words representation on the frequency of the types of all objects in the image. On the other hand, the *GraphWalk* approach obtains multiple subsets of the objects via graph walking and constructs multiple bag-of-words representations from these object subsets. In this thesis, we define the second approach to understand the

Table 4.4: Test set results obtained by our proposed algorithms *Simple* approach and *GraphWalk* approach and the structural comparison algorithms

	Normal	Low-Grade	High-Grade	Overall
Simple Approach (LOPs)	95.32	92.54	90.27	93.03
GraphWalk Approach (nLOPs)	93.68	91.23	90.66	91.90
Delaunay Triangulation	89.61	71.56	87.55	79.71
Color Graph	92.67	82.46	86.38	86.24
Hyrid Model	96.95	88.27	96.11	92.21

Table 4.5: For the test set, the confusion matrix obtained by our *Simple* approach

		Computed		
		Normal	Low	High
Actual	Normal	468	19	4
	Low	16	781	47
	High	0	25	232

effects of using the characterizations of local subregions instead of using the entire image characterization. Our experiments show that although this second approach gives better results than most of the comparison algorithms, it gives statistically significantly worse results than our *Simple* Approach (we use McNemars statistical test with a significance level of 0.05). The confusion matrices for these two algorithms are also given in Table 4.5 and Table 4.6, respectively. As also seen in these tables, for both of these two algorithms, most of the confusions occur in between low-grade and high-grade cancerous tissues. This is indeed consistent with the current practice, in which incorrect decisions are typically observed in grading especially when tissues lie at the boundary between low-grade and high-grade cancer.

Table 4.6: For the test set, the confusion matrix obtained by our *GraphWalk* approach

		Computed		
		Normal	Low	High
Actual	Normal	460	25	6
	Low	28	770	46
	High	3	21	233

4.5 Parameter Analysis

Next, we analyze the effects of the model parameters to the classification accuracy. We analyze them separately for our *Simple* approach and *GraphWalk* approach.

4.5.1 Simple Approach

The *Simple* approach has three model parameters: minimum circle radius r_{min} , highest degree m and cluster number k and one external parameter C in a linear kernel SVM classifier. We select the values of these parameters by three fold cross-validation. Next, we analyze the effects of the three model parameters. For that; we fix two parameters and analyze accuracy as a function of the other parameter.

4.5.1.1 Minimum Radius r_{min}

The minimum circle r_{min} is a threshold for an object radius in tissue image decomposition. Smaller values for the minimum circle r_{min} allow a lot of objects to be defined in tissue decomposition so that the representation may contain irrelevant or noisy false objects. Thus, noisy or false neighborhoods can be defined in extracting local object pattern descriptors. Note that when neighborhoods are incorrectly defined, a tissue may incorrectly be modeled, which decreases classification accuracy.

On the contrary, larger values for the minimum circle r_{min} result in ignoring

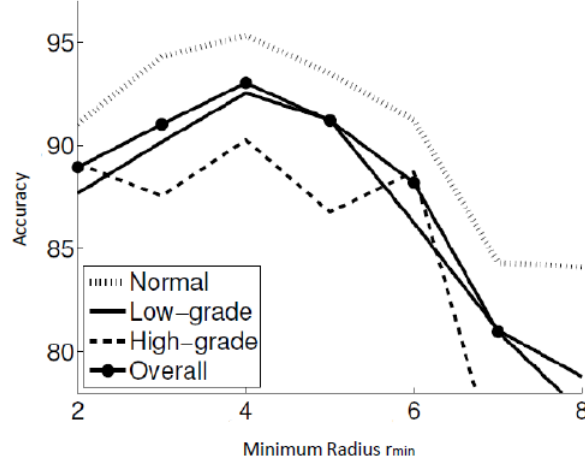


Figure 4.1: Test set accuracy as a function of the minimum circle radius r_{min}

a lot of objects since many objects do not meet the threshold condition. This may cause not to consider completely or partially some important histological components. This might be an important problem especially for nucleus objects as their radii are typically smaller. As a result, this causes incorrect tissue classification. The effects of the minimum circle r_{min} to the accuracy are illustrated in Figure 4.1. This figure also shows the accuracy changes for normal, low-grade, and high grade cancerous tissues.

4.5.1.2 Highest Degree m

We define local object patterns in the 2^k neighborhood for an object where $k = \{1, 2, \dots, m\}$ and model tissues using these patterns. Thus, parameter m is the highest degree in the set of $S = \bigcup_{j=0}^m 2^j$ -LOP, which determines the size of the local object patterns set. Using larger values for the highest degree m increases the number of neighbors in the neighborhood and increases the size of the set of the local object patterns for an object. This results in losing locality in descriptor definition and as a result accuracies decrease. On the other hand, smaller values result in considering only the closest neighbors for pattern extraction. This causes to define non-distinctive descriptors for objects in different classes. As a consequence, this situation makes it difficult to model the tissue

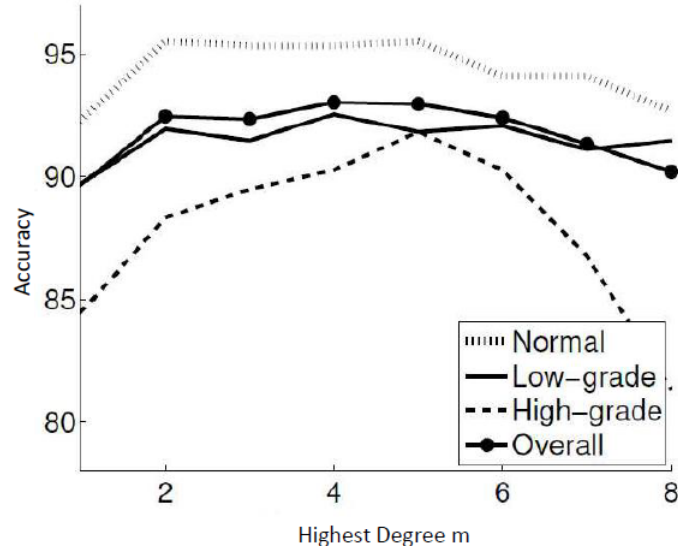


Figure 4.2: Test set accuracy as a function of the highest degree m

classes since cancerous tissues differentiate the normal class for relatively large m values. Non-distinctive descriptors cannot detect class differences so accuracies decrease. Figure 4.2 illustrates the effects of this parameter to classification accuracies.

4.5.1.3 Cluster Number k

In order to define new object types and the visual words of a bag-of-words representation, we separately cluster the objects into k groups. If small values are used for the parameter k , insufficient words are defined, and thus, words are non-distinctive to classify the images. On the other hand, larger values for this parameter increase the number of the defined words. This may decrease the accuracy as a result of curse of dimensionality in classification. Figure 4.3 illustrates the effects of this parameter to accuracies.

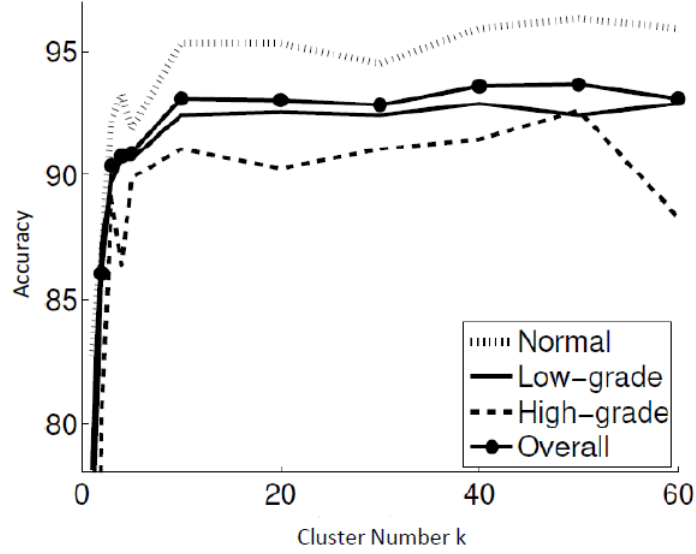


Figure 4.3: Test set accuracy as a function of the cluster number k

4.5.2 GraphWalk Approach

Next, we analyze the parameters of the *GraphWalk* approach. Similarly, we investigate the effects of a single parameter, fixing the remaining ones. Here we observe that the effects of the cluster number k and the highest degree m to the classification accuracy are very similar to those of the *Simple* approach. Thus, we provide the sensitivity analysis for the remaining parameters.

4.5.2.1 Subset Number N

The *GraphWalk* approach obtains N object subsets by graph walking. These subsets do not cover the entire image but correspond to local areas. Using too small values for this parameter causes to ignore some important parts of the image in classification. Thus, it lowers the accuracy. On the other hand, using larger values does not significantly affect the accuracies. Figure 4.4 illustrates the effects of the subset number N to the accuracies.

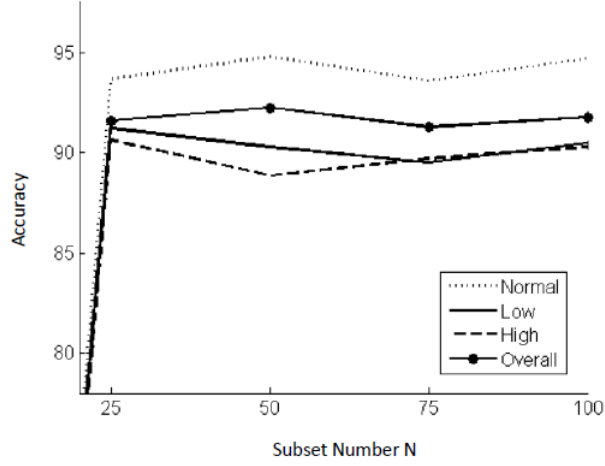


Figure 4.4: Test set accuracy as a function of the subset number N

4.5.2.2 Visited Object Number L

This approach continues graph walking until L objects are visited. Using larger values shows that the use of subregions is not effective as the use of the entire image. This result is consistent with our comparison with the *Simple Approach*. On the other hand, smaller values of L result in not considering some distinctive parts of the tissue image, which lowers the accuracy. Figure 4.5 summarizes the effects of this parameter to the classification accuracy.

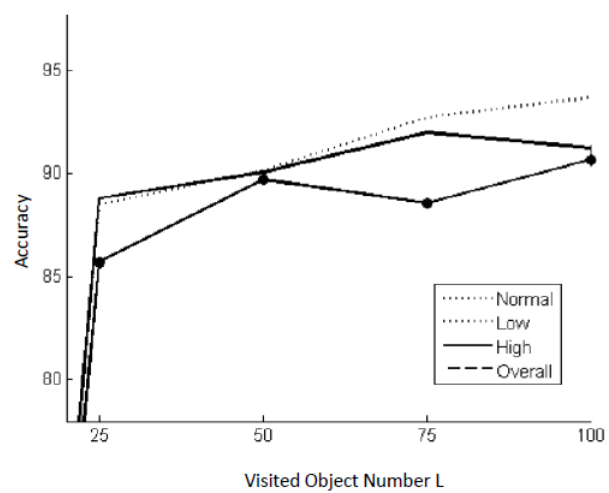


Figure 4.5: Test set accuracy as a function of the visited object number L

Chapter 5

Conclusion

This thesis presents a new algorithm for representing and classifying colon tissue images. In this algorithm, we introduce a set of new high-level texture descriptors called local object patterns. We define these descriptors on tissue objects, which approximately represent histological tissue components. To this end, we specify a set of neighborhoods with different locality ranges and construct a binary string for each of these neighborhoods to encode spatial arrangements of the objects within the specified local neighborhoods. We then characterize tissue objects using the decimal equivalents of the binary strings as descriptors and construct bag-of-words representation of an image from its characterized objects. We implement two algorithms to extract bag-of-word representations: The *Simple* approach uses all objects whereas the *GraphWalk* approach uses multiple object subsets obtained through graph walking. We test our proposed algorithm on 3236 microscopic images of colon tissues stained with hematoxylin and eosin. Our experiments demonstrate that our algorithm, which uses local object pattern descriptors, leads to higher classification accuracies than its pixel-based counterparts.

The proposed algorithm constructs a binary string to encode objects' composition in a specified local neighborhood. In this binary string, purple objects, which usually correspond to nucleus components, are represented with 1 and the

others with 0. Instead of this binary representation, one could consider constructing ternary strings where pink and white objects are represented with different values. Besides, the proposed algorithm computes the local object pattern descriptors by converting the binary strings to their decimal equivalents. It is also possible to obtain these descriptors directly from the strings. In this work, we use the breadth search algorithm in the *GraphWalk* approach. Another future work is to use different graph walks to obtain the object subsets. Exploring these possibilities could be considered as future research directions of this thesis.

Bibliography

- [1] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, “Multifeature prostate cancer diagnosis and gleason grading of histological images,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 10, pp. 1366–1378, 2007.
- [2] R. Rahmadwati, G. Naghdy, M. Ros, and C. Todd, “Computer aided decision support system for cervical cancer classification,” *The Proceedings of SPIE: Applications of Digital Image Processing XXXV*, vol. 8499, pp. 1–13, 2012.
- [3] F. Bunyak, A. Hafiane, and K. Palaniappan, “Histopathology tissue segmentation by combining fuzzy clustering with multiphase vector level sets,” in *Software Tools and Algorithms for Biological Systems*, vol. 696, pp. 413–424, Springer New York, 2011.
- [4] J. S. M. Wiltgen, A. Gerger, “Tissue counter analysis of benign common nevi and malignant melanoma,” *International journal of medical informatics*, vol. 69, no. 1, pp. 17 – 28, 2003.
- [5] A. N. Esgiar, R. N. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, “Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 2, pp. 197–203, Sept. 1998.
- [6] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, “A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1205–1218, 2012.

- [7] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of prostate cancer using architectural and textural image features," *2007 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1284 – 1287, 2007.
- [8] A. Karahaliou, I. Boniatis, S. Skiadopoulos, F. Sakellarpoulos, N. Arikidis, E. Likaki, G. Panayiotakis, and L. Costaridou, "Breast cancer diagnosis: Analyzing texture of tissue surrounding microcalcifications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 731 – 738, 2008.
- [9] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 6, pp. 697–704, 2003.
- [10] 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," in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 496–499, 2008.
- [11] A. Esgiar, R. N. G. Naguib, B. Sharif, M. Bennett, and A. Murray, "Fractal analysis in the detection of colonic cancer images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 1, pp. 54–58, 2002.
- [12] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1037–1050, 2009.
- [13] C. Atupelage, H. Nagahashi, M. Yamaguchi, T. Abe, A. Hashiguchi, and M. Sakamoto, "Multifractal feature descriptor for diagnosing liver and prostate cancers in h&e stained histologic images," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pp. 298–301, 2012.

- [14] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. N. Gurcan, “Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2008, 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part II* (D. N. Metaxas, L. Axel, G. Fichtinger, and G. Szekely, eds.), vol. 5242 of *Lecture Notes in Computer Science*, pp. 196–204, Springer, 2008.
- [15] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, “Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development,” *Pattern Recognition*, vol. 42, pp. 1093–1103, June 2009.
- [16] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, “Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles,” *Machine Vision and Applications*, pp. 1–16, 2012.
- [17] K. Masood and N. Rajpoot, “Texture based classification of hyperspectral colon biopsy samples using clbp,” in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1011–1014, 2009.
- [18] B. Zhang, “Breast cancer diagnosis from biopsy images by serial fusion of random subspace ensembles,” in *4th International Conference on Biomedical Engineering and Informatics (BMEI)*, pp. 180–186, 2011.
- [19] A. Basavanhally, S. Ganesan, S. Agner, J. Monaco, M. Feldman, J. Tomaszewski, G. Bhanot, and A. Madabhushi, “Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 642–653, 2010.
- [20] A. Basavanhally, S. Ganesan, N. Shih, C. Mies, M. Feldman, J. Tomaszewski, and A. Madabhushi, “A boosted classifier for integrating multiple fields of view: Breast cancer grading in histopathology,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 125–128, 2011.

- [21] S. Doyle, M. Feldman, J. Tomaszewski, N. Shih, and A. Madabhushi, “Cascaded multi-class pairwise classifier (cascampa) for normal, cancerous, and cancer confounder classes in prostate histology,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 715–718, 2011.
- [22] H.-K. Choi, T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P.-U. Malmström, and C. Busch, “Image analysis based grading of bladder carcinoma. comparison of object, texture and graph based methods and their reproducibility,” *Analytical Cellular Pathology*, vol. 15, no. 1, pp. 1–18, 1997.
- [23] C. Demir, S. H. Gultekin, and B. Yener, “Learning the topological properties of brain tumors,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 262–270, 2005.
- [24] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir, “Color graphs for automated cancer diagnosis and grading,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 665–674, 2010.
- [25] T. Ojala, M. P. Inen, S. Member, and T. M. A, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [26] R. Siegel, D. Naishadham, and A. Jemal, “Cancer statistics, 2013,” *CA Cancer J Clin*, vol. 63, no. 1, pp. 11 – 30, January/February, 2013.
- [27] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, “Hematoxylin and eosin staining of tissue and cell sections,” *Cold Spring Harbor Protocols*, vol. 2008, no. 5, 2008.
- [28] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, “Histopathological image analysis: a review,” *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147 – 171, 2009.
- [29] I. D. R. M. Haralick, K. Shanmugam, “Texture features for image classification,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 660 – 621, November 1973.

- [30] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172 – 179, 1975.
- [31] M. Bibbo, F. Michelassi, P. H. Bartels, H. Dytch, C. Bania, E. Lerma, and A. G. Montag, "Karyometric marker features in normal-appearing glands adjacent to human colonic adenocarcinoma," *Cancer Research*, vol. 50, no. 1, pp. 147–51, 1990.
- [32] X. Sun, S.-H. Chuang, J. Li, and F. McKenzie, "Automatic diagnosis for prostate cancer using run-length matrix method," *Medical Imaging 2009: Computer-Aided Diagnosis*, vol. 7260, no. 2009, 2009.
- [33] B. Weyn, W. Jacob, V. D. da Silva, R. Montironi, P. W. Hamilton, D. Thompson, H. G. Bartels, A. Van Daele, K. Dillon, and P. H. Bartels, "Data representation and reduction for chromatin texture in nuclei from premalignant prostatic, esophageal, and colonic lesions," *Cytometry*, vol. 41, no. 2, pp. 133–8, 2000.
- [34] M. Sharma and S. Singh, "Evaluation of texture methods for image analysis," in *Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001*, pp. 117–121, 2001.
- [35] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognition Letter*, vol. 11, pp. 415 – 420, June 1990.
- [36] B. R. Dasarathy and E. B. Holder, "Image characterizations based on joint gray-level run-length distributions," *Pattern Recognition Letter*, vol. 12, pp. 497 – 502, 1991.
- [37] K. I. Laws, *Textured Image Segmentation*. PhD thesis, Univ. Southern California, 1980.
- [38] M. Datar, D. Padfield, and H. Cline, "Color and texture based segmentation of molecular pathology images using hsoms," in *2008 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 292–295, 2008.

- [39] N. Sarkar and B. Chaudhuri, “An efficient differential box-counting approach to compute fractal dimension of image,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 1, pp. 115–120, 1994.
- [40] T. Ojala, M. Pietikinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.
- [41] D. T. Lee and B. J. Schachter, “Two algorithms for constructing a delaunay triangulation,” *International Journal of Computer and Information Sciences*, vol. 9, no. 3, pp. 219 – 242, 1980.
- [42] P. Khurd, L. Grady, A. Kamen, S. Gibbs-Strauss, E. Genega, and J. V. Frangioni, “Network cycle features: Application to computer-aided gleason grading of prostate cancer histopathological images,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1632–1636, IEEE, 2011.
- [43] C. Lopez, S. Agaian, I. Sanchez, A. Almunashri, O. Zinalabdin, A. Rikabi, and I. Thompson, “Exploration of efficacy of gland morphology and architectural features in prostate cancer gleason grading,” in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2849–2854, 2012.
- [44] G. Landini and I. E. Othman, “Architectural analysis of oral cancer, dysplastic, and normal epithelia,” *Cytometry Part A*, vol. 61A, no. 1, pp. 45–55, 2004.
- [45] A. C. Ruifrok and D. A. Johnston, “Quantification of histochemical staining by color deconvolution,” *Analytical and quantitative cytology and histology, the International Academy of Cytology and American Society of Cytology*, vol. 23, no. 4, pp. 291–299, 2001.
- [46] A. Tosun, M. Kandemir, C. Sokmensuer, and C. Gunduz Demir, “Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection,” *Pattern Recognition*, vol. 42, pp. 1104–1112, 2009.

- [47] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at url-<http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [48] E. Ozdemir, C. Sokmensuer, and C. Gunduz-Demir, “A resampling-based markovian model for automated colon cancer diagnosis,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 1, pp. 281–9, 2012.
- [49] E. Ozdemir and C. Gunduz-Demir, “A hybrid classification model for digital pathology using structural and statistical pattern recognition,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 474–483, 2013.