HISTOPATHOLOGICAL IMAGE CLASSIFICATION USING SALIENT POINT PATTERNS

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

Celal Çığır

August, 2011

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. Çiğdem 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. İbrahim Körpeoğlu

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. Özlen Konu

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

HISTOPATHOLOGICAL IMAGE CLASSIFICATION USING SALIENT POINT PATTERNS

Celal Çığır
M.S. in Computer Engineering
Supervisor: Assist. Prof. Dr. Çiğdem Gündüz Demir
August, 2011

Over the last decade, computer aided diagnosis (CAD) systems have gained great importance to help pathologists improve the interpretation of histopathological tissue images for cancer detection. These systems offer valuable opportunities to reduce and eliminate the inter- and intra-observer variations in diagnosis, which is very common in the current practice of histopathological examination. Many studies have been dedicated to develop such systems for cancer diagnosis and grading, especially based on textural and structural tissue image analysis. Although the recent textural and structural approaches yield promising results for different types of tissues, they are still unable to make use of the potential biological information carried by different tissue components. However, these tissue components help better represent a tissue, and hence, they help better quantify the tissue changes caused by cancer.

This thesis introduces a new textural approach, called Salient Point Patterns (SPP), for the utilization of tissue components in order to represent colon biopsy images. This textural approach first defines a set of salient points that correspond to nuclear, stromal, and luminal components of a colon tissue. Then, it extracts some features around these salient points to quantify the images. Finally, it classifies the tissue samples by using the extracted features. Working with 3236 colon biopsy samples that are taken from 258 different patients, our experiments demonstrate that Salient Point Patterns approach improves the classification accuracy, compared to its counterparts, which do not make use of tissue components in defining their texture descriptors. These experiments also show that different set of features can be used within the SPP approach for better representation of a tissue image.

Keywords: Salient point patterns, texture, histopathological image analysis, automated cancer diagnosis and grading, colon cancer.

ÖZET

ÖZELLİKLİ NOKTA MODELLERİ KULLANARAK HİSTOPATOLOJİK RESİMLERİN SINIFLANDIRILMASI

Celal Çığır
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Y. Doç Dr. Çiğdem Gündüz Demir
Ağustos, 2011

Son on yıl içinde, bilgisayar destekli teşhis sistemleri, patologların kanser tespiti için histopatolojik görüntüleri yorumlamasını artırmaya yardımcı olması yönüyle büyük bir önem kazanmıştır. Bu sistemler, kanser tanısı için mevcut histopatolojik doku muayenesi uygulamasında çok yaygın olan gözlemci-içi ve gözlemciler arası değişkenliği azaltmaya ve ortadan kaldırmaya yönelik çok değerli fırsatlar sunmaktadır. Özellikle dokusal ve yapısal doku görüntü analizine dayalı birçok çalışma, kanserin tanı ve sınıflandırması için bu tür sistemleri geliştirmeye adanmıştır. Son zamanlardaki dokusal ve yapısal yaklaşımlar, farklı tipte dokular için umut verici sonuçlar vermesine rağmen, doku bileşenleri tarafından taşınan potansiyel biyolojik bilgiyi kullanabilmekten yoksundurlar. Halbuki, bu doku bileşenleri, doku temsiline ve dolayısıyla, kanserin yol açtığı doku değişikliklerini ölçmeye daha iyi yardımcı olur.

Bu tez, kolon biyopsi görüntülerini temsil etmede doku bileşenlerinin kullanımı için Özellikli Nokta Modelleri olarak adlandırılan yeni bir dokusal yaklaşım sunmaktadır. Bu dokusal yaklaşım öncelikle kolon dokusunun çekirdek, stroma ve lümen bileşenlerine karşılık gelen bir dizi özellikli noktaları tanımlar. Sonra, bu belirgin noktalar etrafından doku görüntülerini ölçmede kullanılan öznitelikler çıkartılır. Son olarak, bu öznitelikleri kullanarak doku örneklerini sınıflandırır. 258 farklı hastadan alınan 3236 kolon biyopsi örneği üzerinde gerçekleştirdiğimiz deneyler, Özellikli Nokta Modelleri yaklaşımının, dokuları tanımlamada yapısal bileşenleri kullanmayan benzer çalşmalarla karşılaştırıldığında, sınıflandırma başarı yüzdesini artırdığını ortaya koymuştur. Ayrıca gerçekleştirdiğimiz bu deneyler, doku görüntüsünün daha iyi temsil edilebilmesi için bu dokusal yaklaşım kullanılarak farklı özniteliklerin elde edilebileceğini göstermektedir.

Anahtar sözcükler: Özellikli nokta modelleri, yapısal doku, histopatolojik görüntü analizi, otomatik kanser tanısı ve derecelendirilmesi, kolon kanseri.

Acknowledgement

This thesis would not have been possible without the guidance and the help of several people who contributed and extend their valuable assistance and support in the preparation and completion of this study.

First and foremost, I cannot find words to express my utmost gratitude to Assist. Prof. Dr. Çiğdem Gündüz Demir whose sincerity, encouragement, and continuous support I will never forget. My special thanks goes to my thesis committee members, Assoc. Prof. Dr. İbrahim Körpeoğlu and Assist. Prof. Dr. Özlen Konu, who have graciously agreed to serve on my committee. I would also like to thank Prof. Dr. Cenk Sökmensüer for his consultancy on medical knowledge and Assist. Prof. Dr. Selim Aksoy for teaching Image Analysis course. Thanks to TÜBİTAK-BİDEB and TÜBİTAK-İLTAREN for their financial and research supports to me.

There are some people I would like to thank individually. I want to thank Yaşar Kemal Alp for being such a wonderful friend. It was a fabulous and extremely quiet experience to be sharing a dormitory room with him for five years. I am extremely grateful to my friends for their help and encouragement: Mücahid Kutlu, Bahri Türel, Esat Belviranlı, Alptuğ and Merve Dilek, Cem Aksoy, Ömer Faruk Uzar, Akif Burak Tosun, Erdem Özdemir, Hamza Soğancı, Anıl Bayram, and Serdar Akbayrak. I would also like to acknowledge Gülden Olgun, Salim Arslan, and Can Koyuncu for allowing me to use their computers.

Last but not the least, I would like to thank my parents, Selver and Habip Çığır, and Aynur, my future wife, for their endless support and love. I will be forever grateful to them.

This thesis is dedicated to people who devoted themselves to my country.

Celal Çığır
August, 2011

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer, also named as malignant neoplasm, is the name for a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. Around the world, cancer is ranked the third leading cause of deaths following cardiovascular and infectious diseases. In respect of percentage, such diseases caused almost 13.4 percent of all deaths in men and 11.8 percent in women in 2004 [6]. As it is presented in Figure 1.1, cancer causes more deaths than respiratory diseases, diabetic diseases, and deaths because of perinatal conditions; furthermore cancer results in more deaths than HIV/AIDS, tuberculosis, and malaria which lie in infectious diseases.

Normally, body cells grow over time and new cells take place when old ones die. However, cancer cells grow and divide without dying and form new abnormal cancer cells. At the end, these cells group together and form an additional mass of tissue. This mass is called a malignant tumor. Prostate, breast, lung, and colorectal cancers are some of the cancer types that form a malignant tumor. Some cancers, like leukemia, do not form tumors. Instead, these cancer cells divide irregularly causing increase in white blood cells [1].

Being a tumor-forming cancer type, colon cancer is one of the most common types of cancer that afflicts many people each year. It is also called as colorectal cancer or large bowel cancer. According to American Cancer Society research in

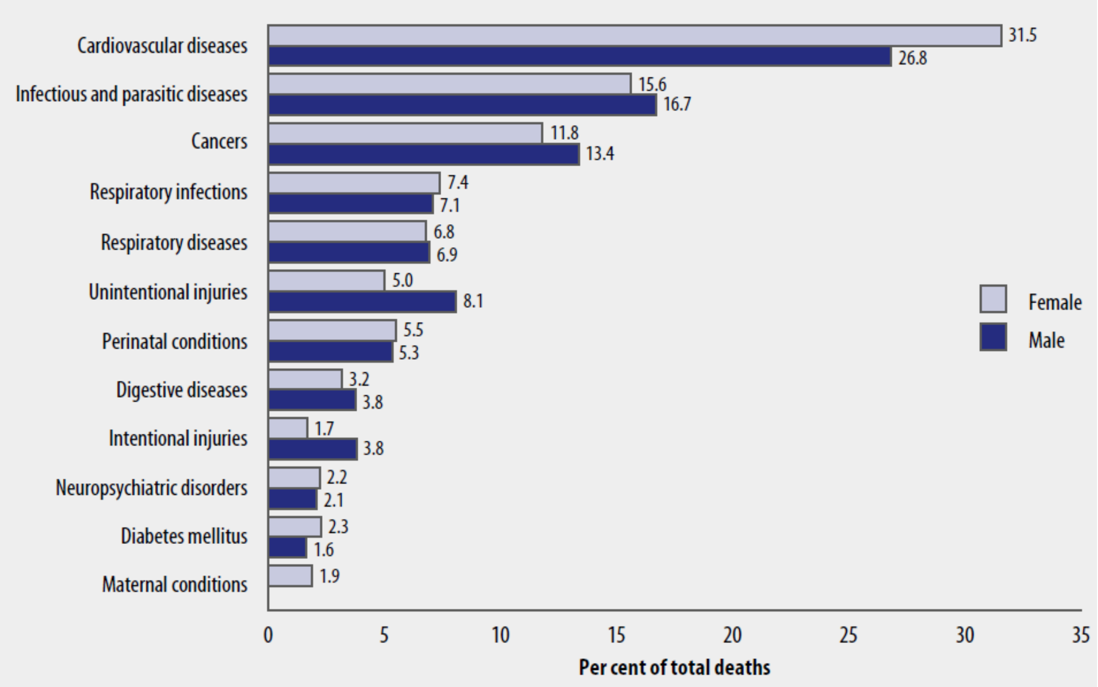
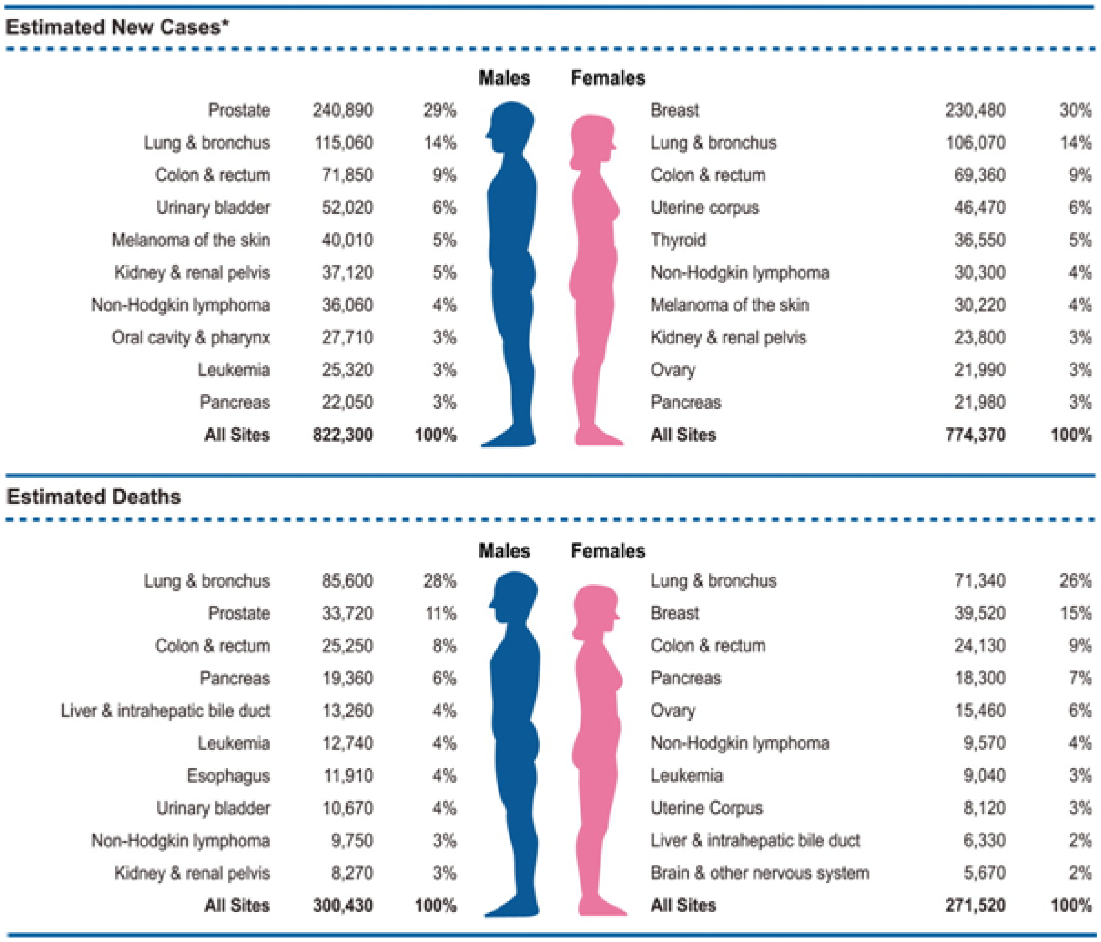


Figure 1.1: Distribution of deaths by leading cause groups, males and females, worldwide, 2004.[1]

2011, about 1.5 million new cancer cases will occur in the U.S. and colon cancer is estimated as the third most common cancer type for both males and females [63]. As it is shown in Figure 1.2, 9 percent of new cancer cases will be type of colorectal cancer and unfortunately, 34 percent of them will result in death for both males and females.

Colon cancer grows in the wall of the colon. Most begin as a small growth on the bowel wall. These growths are usually benign (not cancerous), but some develop into cancer over time. The process of forming tumor in the colon wall can take many years, which allows time for early detection with screening tests. Widespread screening tests play an important role to identify cancers at an early and potentially treatable stage [64]. In order to select the correct treatment plan, cancer must be diagnosed and graded accurately. For cancer diagnosis, there are many methods that are employed in clinical institutions. Blood and urine tests are one of these methods to make cancer diagnosis and they give

[1] **Source**: The Global Burden of Disease: 2004 Update by World Health Organization.



Estimated New Cases*

Males			Females		
Prostate	240,890	29%	Breast	230,480	30%
Lung & bronchus	115,060	14%	Lung & bronchus	106,070	14%
Colon & rectum	71,850	9%	Colon & rectum	69,360	9%
Urinary bladder	52,020	6%	Uterine corpus	46,470	6%
Melanoma of the skin	40,010	5%	Thyroid	36,550	5%
Kidney & renal pelvis	37,120	5%	Non-Hodgkin lymphoma	30,300	4%
Non-Hodgkin lymphoma	36,060	4%	Melanoma of the skin	30,220	4%
Oral cavity & pharynx	27,710	3%	Kidney & renal pelvis	23,800	3%
Leukemia	25,320	3%	Ovary	21,990	3%
Pancreas	22,050	3%	Pancreas	21,980	3%
All Sites	**822,300**	**100%**	**All Sites**	**774,370**	**100%**

Estimated Deaths

Males			Females		
Lung & bronchus	85,600	28%	Lung & bronchus	71,340	26%
Prostate	33,720	11%	Breast	39,520	15%
Colon & rectum	25,250	8%	Colon & rectum	24,130	9%
Pancreas	19,360	6%	Pancreas	18,300	7%
Liver & intrahepatic bile duct	13,260	4%	Ovary	15,460	6%
Leukemia	12,740	4%	Non-Hodgkin lymphoma	9,570	4%
Esophagus	11,910	4%	Leukemia	9,040	3%
Urinary bladder	10,670	4%	Uterine Corpus	8,120	3%
Non-Hodgkin lymphoma	9,750	3%	Liver & intrahepatic bile duct	6,330	2%
Kidney & renal pelvis	8,270	3%	Brain & other nervous system	5,670	2%
All Sites	**300,430**	**100%**	**All Sites**	**271,520**	**100%**

Figure 1.2: Ten leading cancer types for estimated new cancer cases and deaths, by sex, United States, 2011.[2]

the pathologists beneficial information about the effects of the disease on the body. Medical imaging techniques, such as X-ray images, magnetic resonance imaging (MRI), and ultrasonography are used to detect different types of cancer [4, 11, 26, 31, 64, 65]. Genetic testing is another technique for cancer diagnosis but due to complexity of testing, cost, and a requirement for specialists make it hard to apply for practical use in clinical institutions [17, 52].

Although all these screening methods are used to detect cancer, the final diagnosis are made with histopathological tissue examination. Moreover, these aforementioned screening methods are not capable of making reliable grading,

[2] **Source**: Siegel R. *et al.* Cancer statistics, 2011. CA: A Cancer Journal for Clinicians, 61(4):212236, 2011.

and hence, histopathological tissue examination should be done [9, 12, 58]. In clinical medicine, histopathology refers to the examination of a biopsy (a sample tissue) or a surgical specimen by a pathologist, after histological sections have been stained with a special technique to enhance contrast in the microscopic image and placed onto glass slides. In this examination, pathologists visually examine the changes in cell morphology and tissue distribution under a microscope. If a cancerous region is found in a tissue sample, a grade is assigned to the tissue to characterize the degree of its malignancy.

Conventional histopathological examination in cancer diagnosis is prone to subjectivity and may lead to a considerable amount of intra- and inter-observer variation and poor reproducibility, due to its heavy reliance on human interpretation [3, 33, 46]. Moreover, it is a time-consuming process to examine the whole specimen for making a decision. Computer-aided diagnosis (CAD), therefore, has been proposed to eliminate variations among pathologists and decrease the subjectivity level by assisting pathologists with making more reliable decisions in a time-saving way.

1.1 Motivation

In literature, there have been many studies to develop CAD systems to assist pathologists in their evaluations of histopathological images. In these studies, a tissue is represented with a set of mathematical features that is used in automated diagnosis and grading process. In order to extract these features, there are mainly four different approaches. These are intensity-based, textural, tissue component-based (morphological), and structural approaches.

In the intensity-based approach, a tissue is quantified with the statistical distributions of gray level or color intensities of its pixels. This approach first computes a color or gray level histogram of a tissue image by quantizing its pixels into bins and then, defines a set of statistical features on the histogram [11, 65, 67]. However, intensity distributions are similar for different types of tissues stained with the hematoxylin-and-eosin staining technique. Moreover, this approach is not capable of capturing spatial relations of tissue components as visual descriptors of the images.

In the textural approach, tissue images are represented with a set of features that can be calculated from co-occurrence matrices [16, 29, 19, 60], run-length matrices [44, 74], multiwavelet coefficients [32, 74], fractal geometry [20, 72], and local binary patterns (LBP) [41, 61, 65]. Since texture definition is made on pixels, it is sensitive to noise in the pixel values.

In the morphological approach, a tissue is represented with the size, shape, orientation, and other geometric properties of the tissue components. These properties are measured defining morphological features such as area, perimeter, roundness, and symmetry [72, 67]. This approach requires identifying exact boundaries of cells before extracting the features. Due to complex nature of histopathological images, it is hard to locate tissue components and this leads to a difficult segmentation problem.

Structural approach characterizes the tissue with the spatial distribution of its cellular components. A tissue is represented as a graph and a set of structural features is extracted from this graph representation. The locations of the cell nuclei are considered as nodes to generate such graphs including Delaunay triangulations and their corresponding Voronoi diagrams [18, 74], minimum spanning trees [75], and probabilistic graphs [15, 27].

Although the recent textural and structural approaches for the development of CAD systems yield promising results for different types of tissues, they are still unable to make use of the potential biological information carried by the tissue components. However, these tissue components help better represent a tissue, and hence, they help better quantify the tissue changes due to existence of cancer. For example, in typical colon tissues, epithelial cells are arranged in an order around a luminal structure to form a glandular structure and non-epithelial cells take place in stroma found in between these glands. The gland structures for normal colon tissues are presented in Figures 1.3(a) and 1.3(b). This gland formation deviates from its regular structure due to existence of cancer. At

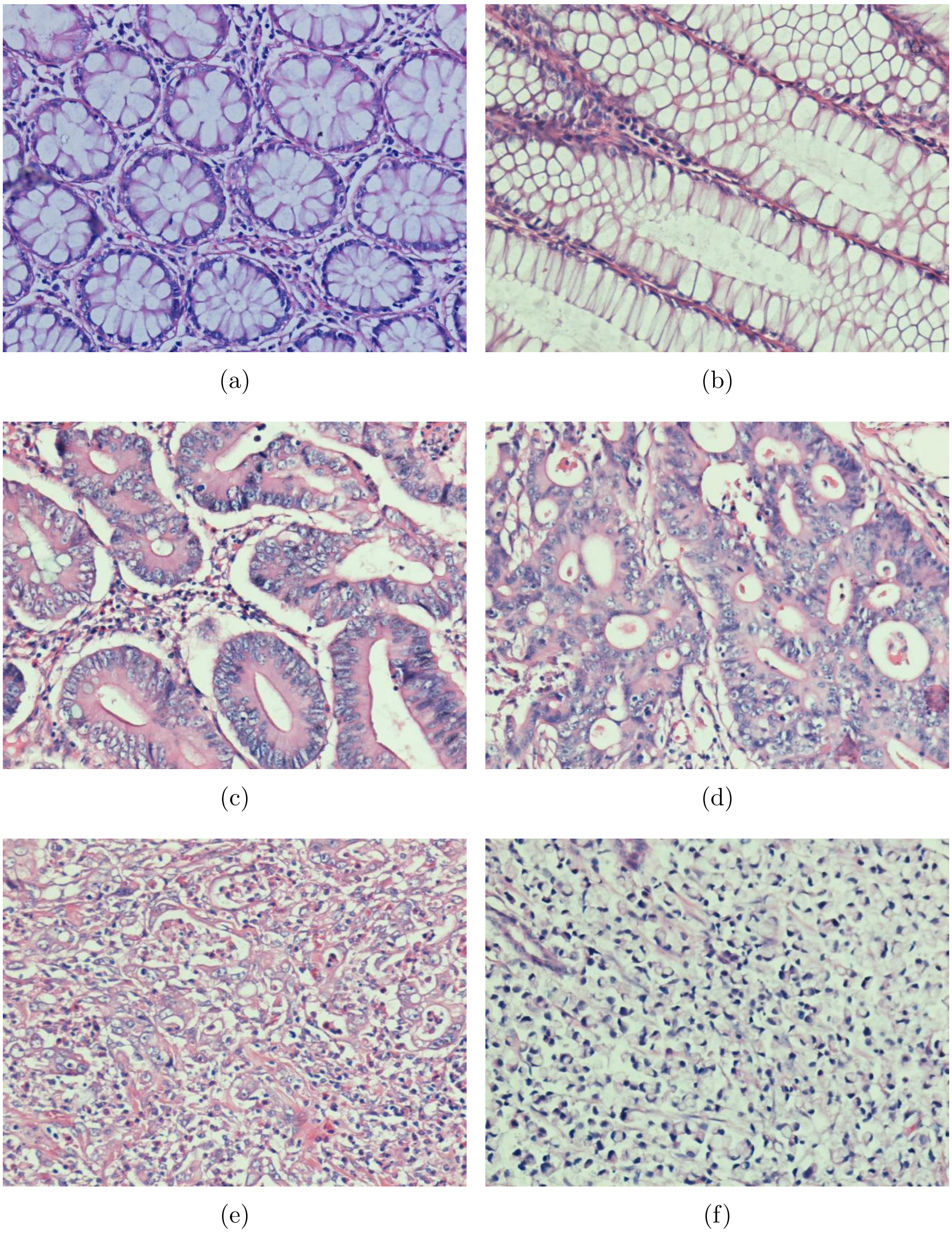


Figure 1.3: Histopathological images of colon tissues, which are stained with the routinely used hematoxylin-and-eosin technique: (a)-(b) normal, (c)-(d) low-grade cancerous, and (e)-(f) high-grade cancerous.

the beginning, the degree of distortion is lower such that gland formations are well to moderately differentiated; examples of low-grade cancerous tissues are shown in Figures 1.3(c) and 1.3(d). Then, the distortion level becomes higher such that the gland formations are poorly differentiated. Figures 1.3(e) and 1.3(f) present such high-grade cancerous colon tissue samples. The quantification of these deviations in glandular structure is very important for accurate cancer grading. In addition to its nuclear components, luminal and stromal regions make easier the quantification of the distortions.

1.2 Contribution

Pathologists visually examine the spatial relations of tissue components such as stroma, nuclei, and lumen for cancer diagnosis and grading. However, most of the textural and structural approaches use only the information provided by nuclear region, ignoring the potential information provided by luminal or stromal regions. On the other hand, it is beneficial to use these tissue components all together for better characterization of a tissue.

In this thesis, we introduce a new textural method, called Salient Point Patterns (SPP), for the utilization of tissue components in order to represent colon biopsy images. For this purpose, first a set of salient points is defined to approximately represent the tissue components including nuclei, stroma, and lumen. Then, a circular window centered on the centroids of the components is used as a mask to extract different types of features around these salient points. Finally, tissue classification is performed by using the extracted set of features. Our experiments demonstrate that this classification approach leads promising results for differentiating normal, low-grade cancerous, and high-grade cancerous tissue images. The main contribution of this thesis can be summarized as the use of salient points to represent tissue components in texture definition. Moreover, the SPP method has a potential of being applied for other types of cancer such as prostate and skin cancer.

1.3 Organization of the Thesis

This thesis is organized as follows: In Chapter 2, we give an overview of the medical background information about cancer and the earlier research related to classification of histopathological images. In Chapter 3, we explain the proposed Salient Point Patterns method in detail. Consequently, in Chapter 4, we describe the experiments and analyze the experimental results. Finally, we summarize our work and discuss its future research aspects in Chapter 5.

Chapter 2

Background

This chapter presents the background information about histopathological image analysis. First of all, general information about colon tissues, the staining process, and changes in colon tissues caused by colon cancer are mentioned. Following the medical background, a brief summary of the previous studies about histopathological image classification for cancer grading and diagnosis is presented. Finally, SIFT key point descriptor, a method for detecting interest points in tissue images, is mentioned.

2.1 Medical Background

The colon, sometimes referred to as the large intestine or large bowel, is a long, hollow tube at the end of the digestive tract. Its main role can be defined as a waste processor; taking digested food in the form of solid waste and pushing it out of the body through the rectum and anus. It absorbs water, electrolytes and nutrients from food and transports them into the bloodstream.

If a doctor suspects cancer in colon, he or she needs to know the status of the colon tissues. In a current clinical practice, histopathological examination is the routinely applied method for diagnosis and grading of colon cancer. Histopathological examination of tissues starts with removing a sample of a small amount of tissue. This procedure is called biopsy. The sample tissue or surgical specimen is to be examined under a microscope by a pathologist. Before the microscopic examination, histological sections have been taken from the biopsy and stained with special chemicals in order to enhance contrast in the microscopic image.

Hematoxylin-and-eosin (H&E) is the routinely used technique for staining tissues at clinical institutions. In a typical tissue, hematoxylin colors nuclei of cells with blue-purple hue, whereas alcoholic solution of eosin colors eosinophilic structures such as proteins and cytoplasms with pink [22]. Therefore, a biopsy tissue stained with the hematoxylin-and-eosin technique has large amounts of different levels of blue-purple, pink, and white components. In Figure 2.1, a sample colon tissue stained with the hematoxylin-and-eosin technique is presented. In a typical colon tissue, there are epithelial cells and stromal cells. An epithelial cell consists of a nucleus, dark purple region, and a cytoplasm, white region near the epithelial cell nucleus. A sample epithelium cell is marked with a red solid circle in the figure. A group of epithelium cells is lined up around a luminal region to form a glandular structure; a luminal area and a gland border are also marked in this figure. Stromal cells are not part of the glandular structures. They are connective tissue components that hold all of the structures in the tissue together.

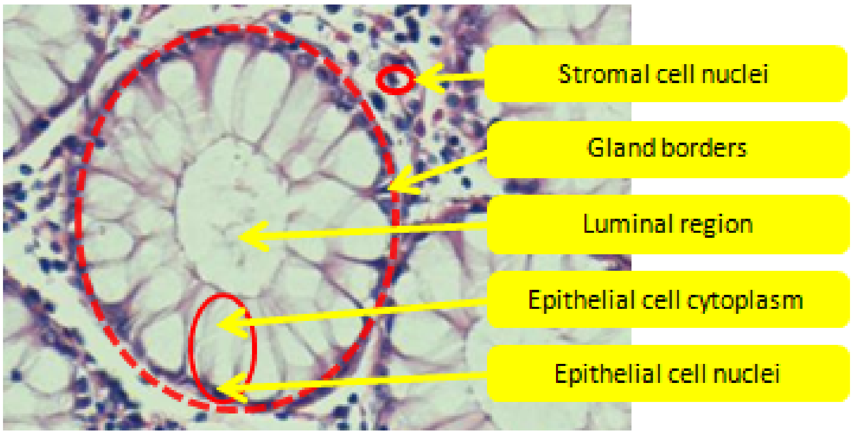


Figure 2.1: Cellular, stromal, and luminal components of a colon tissue stained with the hematoxylin-and-eosin technique.

Colon adenocarcinoma, which accounts for 90-95 percent of all colorectal cancers, originates from the lining of the large intestine causing organizational changes in the glandular structure of the colon tissue. In order to determine the most appropriate treatment plan, cancer should be diagnosed and graded accurately. For colon adenocarcinoma, grading is a description of how closely a colorectal gland looks like a normal gland. It scales the distortions in the organization of the colon tissue. In low-grade cancerous tissues, gland formations are well to moderately differentiated. However, in high-grade cancerous tissues, gland formations are only poorly differentiated [21]. In this thesis, we consider classifying a tissue image as normal, low-grade cancerous, or high-grade cancerous.

2.2 Automated Histopathological Image Analysis

In literature, there have been several studies related to the development of CAD systems for histopathological image analysis. These studies can be broadly divided into three main groups based on their purpose: tissue image segmentation, retrieval, and classification.

2.2.1 Tissue image segmentation

Segmentation is the initial step in histopathological image analysis. Here, the aim is to divide a heterogenous image into its homogeneous parts. These homogeneous parts can later be used by classification algorithms. To identify the homogenous regions in an image, many approaches have been proposed. Tosun *et al.* [70] proposed a new algorithm for an unsupervised segmentation of colon biopsy images. In their algorithm, they first run k-means clustering on the color intensities of pixels to cluster them into three main groups: purple for epithelial and lymphoid cell nuclei regions, pink for connective tissue regions and epithelial cell cytoplasm and white for luminal structures, connective tissue regions, and epithelial cell cytoplasm. Then, they define circular primitives on the pixels of each cluster and define a descriptor on these primitives that is to be used in a region growing algorithm. In their recent study, they introduced another descriptor that is used for histopathological image segmentation [69]. In this study, graphs are used to quantify the relations of tissue components.

Kong *et al.* applied clustering-based segmentation on whole slide histology images [36]. They first divide a whole slide image of neuroblastoma tumor into tiles and each image tile is segmented into five salient components: nuclei, cytoplasm, neuropil, red blood cells, and background. Segmentation is performed by constructing a feature vector that combines color and entropy information extracted from the RGB and La*b* color spaces and classifying this vector into one of the components.

Image filtering is also proposed for segmentation of colon biopsy images [78]. In this study, a directional 2D filter is applied to an image. Each directional 2D filter detects the chain segment in a particular direction and gland segmentation is performed on these filter responses. Image thresholding on the color intensities of pixels followed by iterative region growing is another approach for segmentation of histopathological images [77]. Naik *et al.* proposed to use a Bayesian classifier on pixel values to detect nuclei, cytoplasm, and lumen components in prostate tissue images [47, 48]. They manually select a set of pixel values representing each of the three classes as the ground truth. Then, pixels in a tissue image are labeled using a Bayesian classifier after applying an empirically determined threshold. Finally, a region that corresponds to a set of connected pixels in the same cluster is considered as a segmented region.

2.2.2 Medical image retrieval

With the development of medical imaging and computer technology, there is an exponential increase in the amount of medical images and this makes the management, maintenance, and retrieval of medical images more difficult. This fact leads researchers to work on medical image retrieval systems [43]. Such systems are designed for accessing the most visually similar images to a given query image from a database of images.

ASSERT [62] is one of the medical image retrieval systems that mainly focus on the computed tomography (CT) images of lung. Image Retrieval in Medical Applications (IRMA) [34] system is designed for the classification of images into anatomical areas. Traditional content-based medical image retrieval is based on low level features such as color, texture, shape, spatial relationships, and mixture of these [43, 79, 83, 85]. On the other hand, recent studies have focused on the semantic content analysis of medical images in order to improve the retrieval performance [8, 68, 80]. Tang *et al.* propose I-Browse system that is specialized for retrieval of gastrointestinal tract images [68]. They manually assign semantic labels to the subimages with the help of histopathologists to form ground truth image patches. Then, a set of Gabor features and gray level mean and deviations of the normalized histogram of these subimages are extracted to construct features in image retrieval process. Caicedo *et al.* propose a semantic content-based image retrieval for histopathology images [8]. They first extract low level features, such as gray/color histogram, edge histogram, and texture histogram features, on images of special skin cancer called basal-cell carcinoma. Then, these low level features are mapped to high-level features that reflect the semantic content of the images with help of pathologists.

The evaluation of content-based medical image retrieval performance is usually done by measuring precision and recall values defined as follows [43]:

$$precision = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ items\ retrieved} \tag{2.1}$$

$$recall = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ relevant\ items} \tag{2.2}$$

2.2.3 Histopathological image classification

Many studies have been proposed to classify histopathological images in order to support pathologists in their evaluations. In these studies, tissue images are represented with a set of mathematical features. In literature, there are mainly four different approaches to extract mathematical features for tissue representation. These are morphological, intensity-based, textural, and structural approaches.

In the morphological approaches, a tissue is represented with the size, shape, orientation, and other geometric properties of its cellular components. However, this approach requires segmentation of tissue components before hand. One of the earliest studies based on morphological features is done by Street *et al.* [66]. They first segment the nucleus components of breast tumor tissues in a semi-automated way. Then, they define morphological features such as radius, perimeter, area, compactness, smoothness, concavity, and symmetry. Masood *et al.* use morphological features to represent cellular components of the colon tissues [42]. Such components are nuclei, cytoplasm, glandular structures, and stromal regions, which are segmented by applying k-means clustering on color intensities of a colon tissue image. Rajpoot *et al.* extract morphological features such as area, eccentricity, average diameter, Euler number, orientation, solidity, major axis length, and minor axis length for colon tissue cell representations [55]. Moreover, morphological features are commonly used in cancer diagnosis [11, 18, 45, 47, 67, 72]. Here are the definitions of some of these morphological features :

- *Area*: The number of pixels in the region.
- *Perimeter*: The total length of the region boundaries.
- *Radius*: The radius of a circle with the same area as the region.
- *Compactness*: The measure of the compactness of the region using the formula $perimeter^2/area$.
- *Major axis length*: The longest axis length in the region.

- *Minor axis length*: The length of the axis that is orthogonal to the major axis of the region.

In the intensity-based approach, gray level or color intensities of pixels are used to quantify a tissue image. First, color histogram is computed by putting the pixels of the image into bins and then, features such as mean, standard deviation, skewness, kurtosis, and entropy are defined on this histogram [11, 76, 18]. However, this type of features does not include any information about the spatial distributions of tissue components or pixels.

In the textural approach, the texture of an entire tissue or tissue components is quantified by computing different textural features derived from co-occurrence matrices, run-length matrices, Gabor filters, and fractal dimension analysis. Co-occurrence matrices are widely used tools to extract textural descriptors for histopathological image analysis [16, 18, 19, 29, 60, 72]. Doyle *et al.* use co-occurrence matrices and Gabor filter responses to represent prostate tissue images for cancer grading [18]. Waheed *et al.* performe textural analysis on renal cell carcinoma by computing fractal dimension features together with co-occurrence features on an entire image as well as on an individual cellular structures [72]. Esgiar *et al.* study on the textural analysis of cancerous colonic mucosa [19]. They first compute a co-occurrence matrix on gray-level tissue images. Then, angular second moment to characterize the homogeneity of the image, difference moment to measure local variation, correlation function to calculate the linearity of the gray-level dependencies, entropy to measure randomness, inverse difference moment to identify the local homogeneity of the image and dissimilarity to measure the degree of dissimilarity between pixels are computed on the co-occurrence matrix.

In the structural approach, a tissue is represented with the spatial distributions of its cellular components. Graph representation is made on the tissue components and a set of structural features is derived on this graph. Doyle *et al.* employed Delanuay triangulations, minimum spanning trees, and Voronoi diagrams to describe spatial arrangement of the nuclei on prostate tissue images [18]. Features are derived from these graphs including area, the disorder of the area, and roundness factor on a Voronoi diagrams together with the average edge length and maximum edge length on Delanuay triangulations. Altunbay *et al.* propose color graphs for representing colon tissue images [2]. They initially segment nuclear, stromal, and luminal regions from a tissue image and identify their centroids as graph nodes. Then, a Delanuay triangulation is constructed on these nodes by assigning different colors to the edges depending on the component types of their end nodes. Finally, the colored-version of the average degree, average clustering coefficient, and diameter are extracted to quantify the graph, and thus, to represent the tissue.

Moreover, recent studies on histopathological images focus on the local salient points that contain more information about the underlying medical structure. Raza *et al.* propose a CAD system for classification of renal cell carcinoma subtype using the scale invariant feature transform (SIFT) features [57]. In their next study [56], they use the SIFT features to decompose an image into a collection of small patches and then, apply k-means to cluster these small patches. Finally, they represent an image as the number of descriptors assigned to each cluster, called *bag-of-features.* Caicedo *et al.* employ the SIFT keypoints for construction of the *codebook* to represent histopathological image contents [7]. Díaz *et al.* apply the SIFT algorithm to represent local patches that correspond to nuclear structures in skin biopsy images [14].

2.3 SIFT Key Points

Lowe propose a SIFT (Scale Invariant Feature Transform) method for extracting distinctive features that characterize a set of keypoints for an image [39]. The keypoints are shown to be scale and orientation invariant. Therefore, SIFT is used in many studies such as object recognition, image matching applications, and image retrieval applications.

There are four major stages of computation that the SIFT algorithm uses to generate the set of image features :

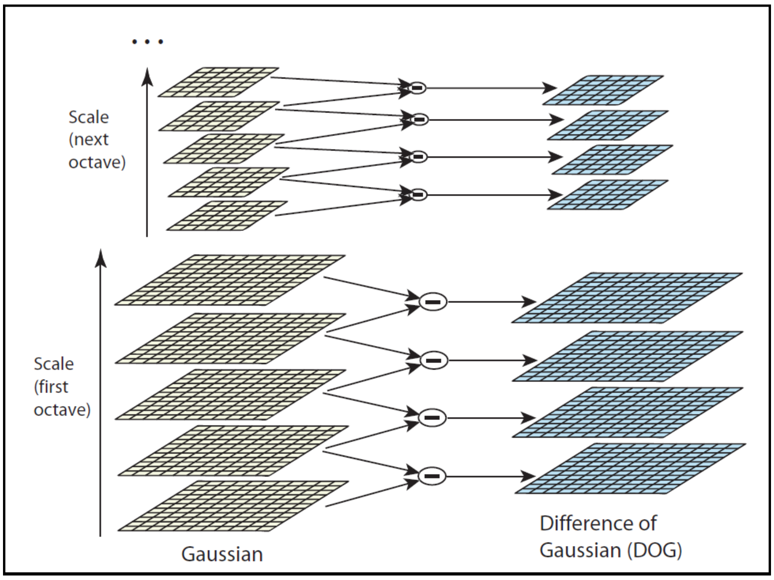


Figure 2.2: The initial image is repeatedly convolved with Gaussians to produce the set of scale space images shown on the left. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images on the right. After each octave (an octave corresponds to doubling the value of σ), the Gaussian image is down-sampled by a factor of 2, and the process repeated.

1. ***Scale-space extreme detection:*** Potential interest points are identified by scanning an image over locations and scales. The keypoints are the local peaks or extreme points in the scale space of the image generated by applying Difference-of-Gaussian (DoG) functions to the image. The scale space of the image is defined as a function, $L(x, y, \sigma)$, that is produced by convolving the input image, $I(x, y)$, with variable scale Gaussian function $G(x, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2.3}$$

where $*$ is the convolution operation and $G(x, y, \sigma)$ is a Gaussian function defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\pi\sigma^2} \tag{2.4}$$

A DOG scale space function $D(x, y, \sigma)$ can be computed from the difference of two scales separated by a constant multiplicative factor of k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{2.5}$$

An efficient approach to construct DOG images is shown in Figure 2.2. First, an initial image I is convolved with a Gaussian function, G_0, of width σ_0. Resulting image, L, is the blurred version of the original image. Then, this blurred image is incrementally convolved with a Gaussian, G_i, of width σ_i to generate the i^{th} image in the stack, which is equivalent to the original image convolved with a Gaussian G_k, of width $k\sigma_0$. Adjacent image scales are subtracted to produce the Difference-of-Gaussian images as shown on the right side of the figure.

2. ***Accurate keypoints localization:*** This stage is to determine a detailed model of location and scale for each candidate location. Least square fitting is conducted via Taylor expansion of the scale-space function, $D(x, y, \sigma)$, so that the origin is at the sample point:

$$D(X) = D + \frac{\partial D^T}{\partial X}X + \frac{1}{2}X^T\frac{\partial^2 D}{\partial X^2}X \tag{2.6}$$

where X $= (x, y, \sigma)^T$ is the offset from this point. Then, keypoints at the location are located and scaled by calculating the extreme of the fitted surface. The keypoints are eliminated if they are found unstable during the computation.

3. ***Orientation assignment:*** Each keypoint location is assigned to several orientations that is based on local image gradient directions in a scale invariant manner. For each image sample, $L(x, y)$, at scale of the keypoint, the gradient magnitude, $m(x, y)$, and orientation, $\theta(x, y)$, is computed using pixel differences:

$$\begin{aligned} m(x,y) &= \sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \\ \theta(x,y) &= \tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y))) \end{aligned}$$

An orientation histogram, which has 36 bins covering the 360 degree, is formed from the gradient orientations. Finally, orientation of the highest magnitude is assigned to the keypoint.

4. ***Keypoints descriptor:*** A descriptor for each keypoint is created. First of all, the coordinates of the descriptor and gradient orientations are rotated

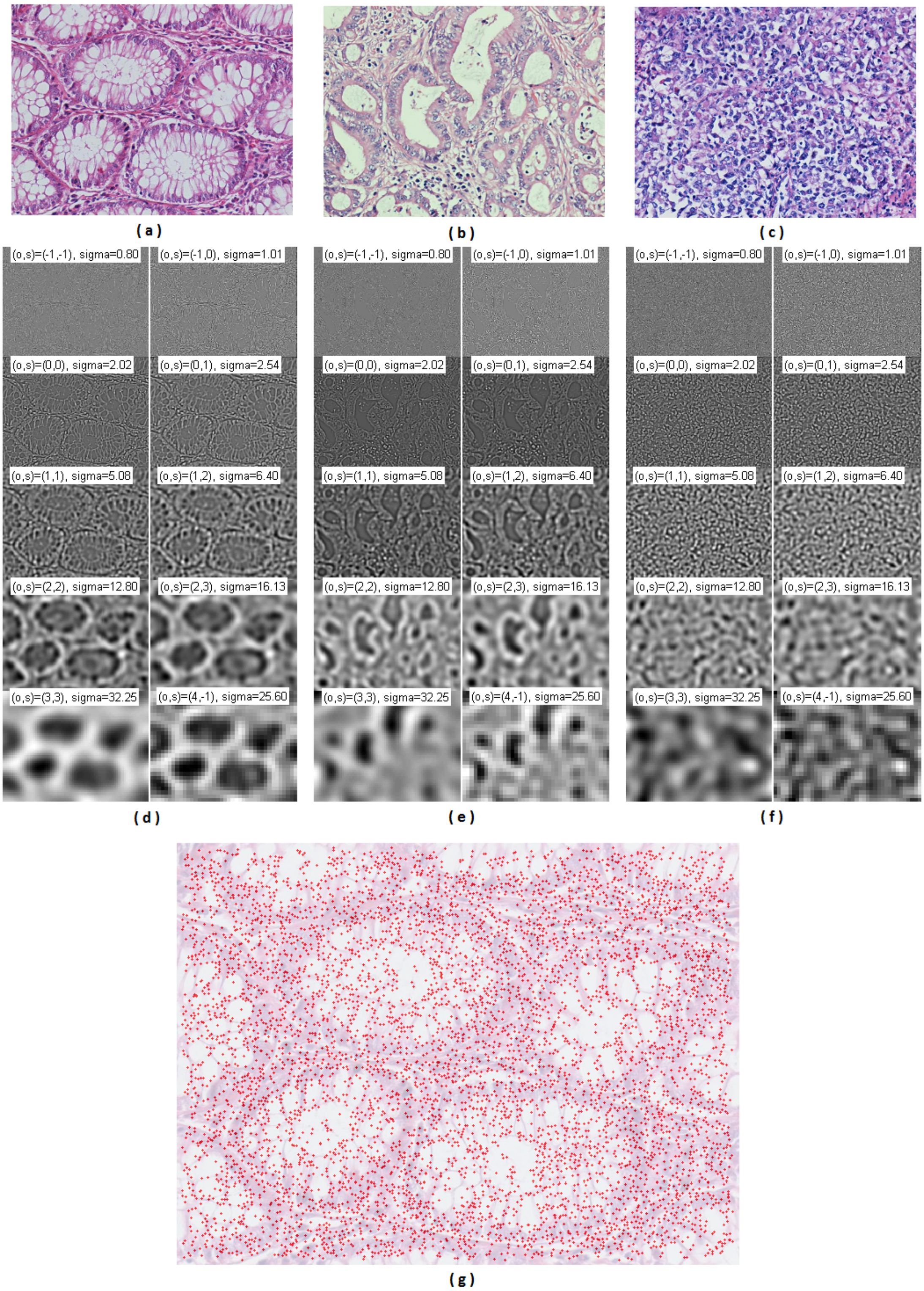


Figure 2.3: DOG images in different scales and octaves: (a) Normal, (b) low-grade cancerous, and (c) high-grade cancerous colon tissues. Their corresponding DOG images in different scales and octaves are (d), (e), and (f), respectively. The SIFT keypoints defined on (a) are presented in (g).

relative to the keypoint orientation in order to achieve rotation invariance. Using the scale of the keypoint, the pixels in the 16×16 neighborhood of the keypoint location are divided into 4×4 windows. The gradient vectors are accumulated into 8 orientation bins resulting in 128 descriptors for each keypoint. Then, the vector is normalized to make it invariant to illumination changes.

An example of DOG images generated from the colon biopsy tissue images can be found in Figure 2.3.

Chapter 3

Methodology

In previous chapters, we mentioned that histopathological examination is prone to subjectivity and may lead to a considerable amount of intra- and inter-observer variation due to its heavy reliance on pathologist interpretation. To eliminate the subjectivity level, and therefore, to help pathologists make more accurate decisions, computer-aided diagnosis (CAD) has been proposed. There have been many studies and methods on the construction of CAD systems as presented in Chapter 2. However, these studies usually discard the domain specific knowledge and treat histopathological images as generic images. In this chapter, we introduce a new method, called Salient Point Patterns (SPP), to characterize the histopathological images with its tissue components: nuclei, stroma and lumen.

The proposed method is composed of a series of processing steps. The first step begins with clustering the pixels of an image into three groups, which correspond to nuclear (purple), stromal (pink), and luminal (white) areas using the k-means clustering algorithm. Then, postprocessing is applied on the pixels of each cluster to decrease the effects of noise due to incorrect clustering of pixels. The next step fits circular structures into these white, purple, and pink areas with the help of the circle-fit transform [28]. The centroids of resulting circular objects constitute salient points to the next step. A set of textural and intensity-based features are computed around these salient points by using a circular window. The features of objects of the same component type are aggregated to define the feature set of the

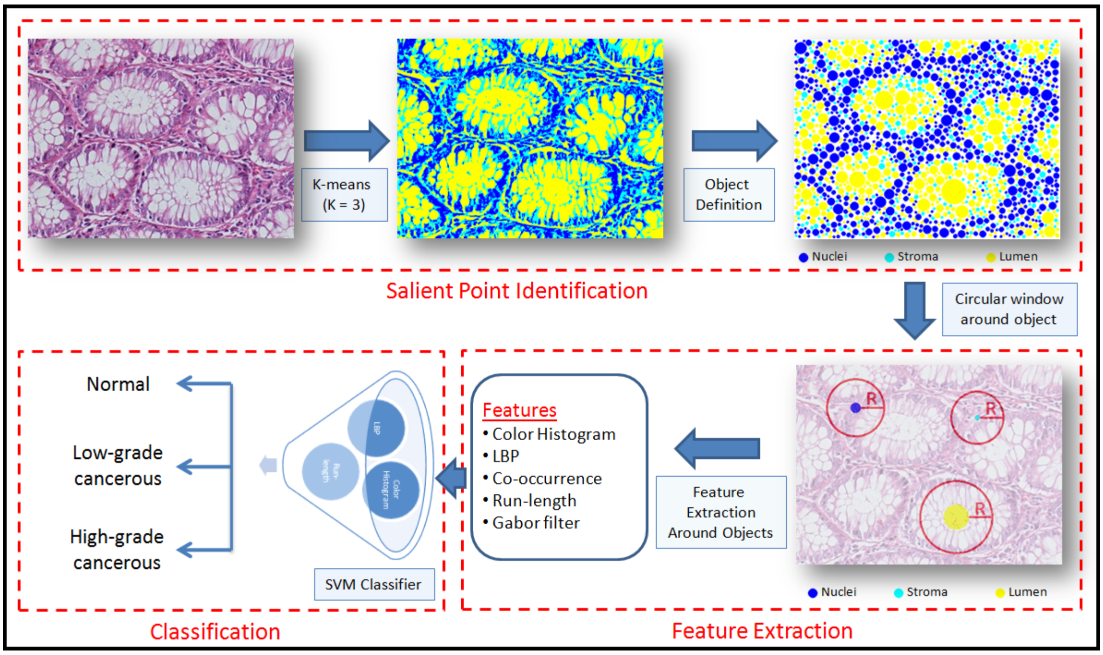


Figure 3.1: The flowchart of the proposed histopathological image processing system.

entire image. Finally, training and classification of the tissue images is performed by using these feature sets. The flowchart of the proposed system is given in Figure 3.1. As shown in this figure, the proposed system consists of three main components: salient point identification, feature extraction, and classification. In this chapter, details of these components are presented.

3.1 Salient Point Identification

Histopathological examination depends on pathologists' visual interpretation of medical images. During this examination, pathologists examine the tissue components and their spatial relations within the tissue images. Therefore, detecting these tissue components may help us define different feature descriptors for colon biopsy images. However, because of the complex nature of a histopathological image scene, it is difficult to exactly segment the components even by a human eye. In a typical histopathological image, there could be staining and sectioning related problems such as existence of touching and overlapping components, heterogeneity of the regions inside a component, and presence of stain artifacts in a tissue [25]. Therefore, instead of determining the exact locations, we approximately describe the tissue components with a set of circular primitives. The centroids of these tissue components are considered as the *salient points*. In this representation, colon tissues stained with hematoxylin-and-eosin (H&E) staining technique have three types of circular primitives: one for nuclear components, one for stromal components, and one for luminal components. The idea behind this approach is inspired from the study presented in [28]. The following sections provide the steps of salient point identification process in detail.

3.1.1 Clustering

In order to segment the tissue components, our system first converts a tissue image from an RGB to La*b* color space. The La*b* color space is developed by the Commission Internatile d'Eclairage (CIE) [51]. It is a perceptually uniform color space when its compared to other color spaces such as RGB, HSI, and YUV. Perceptual uniform means that an amount of change in a color value should result in the same amount of perceptual difference. This allows the use of the Euclidean distance metric in image analysis applications. The La*b* color space is able to represent luminance and chrominance information separately. The L channel carries the information for the light intensity whereas the a* and b* channels represent color intensities.

The k-means clustering algorithm is a process of partitioning or grouping a given N-dimensional set of patterns into k disjoint clusters. This is done such that patterns in the same clusters have similar characteristics and patterns belonging to different clusters have different characteristics. The k-means algorithm has been shown to be effective in producing good clustering results for many applications [81]. The aim of the k-means algorithm is to divide m points in d dimensions into k clusters so that the sum of the squared distance between each point to the centroid of the cluster that it belongs to is minimized where k is the desired number of clusters, C_i with i= 1,2,...,k is the i^{th} cluster containing n_i data points, $0 < n_i < N$, and c_i is the geometric centroid of the cluster C_i.

In other words, the algorithm minimizes the following mean-squared-error cost function for the given N input data points $x_1, x_2, ..., x_N$:

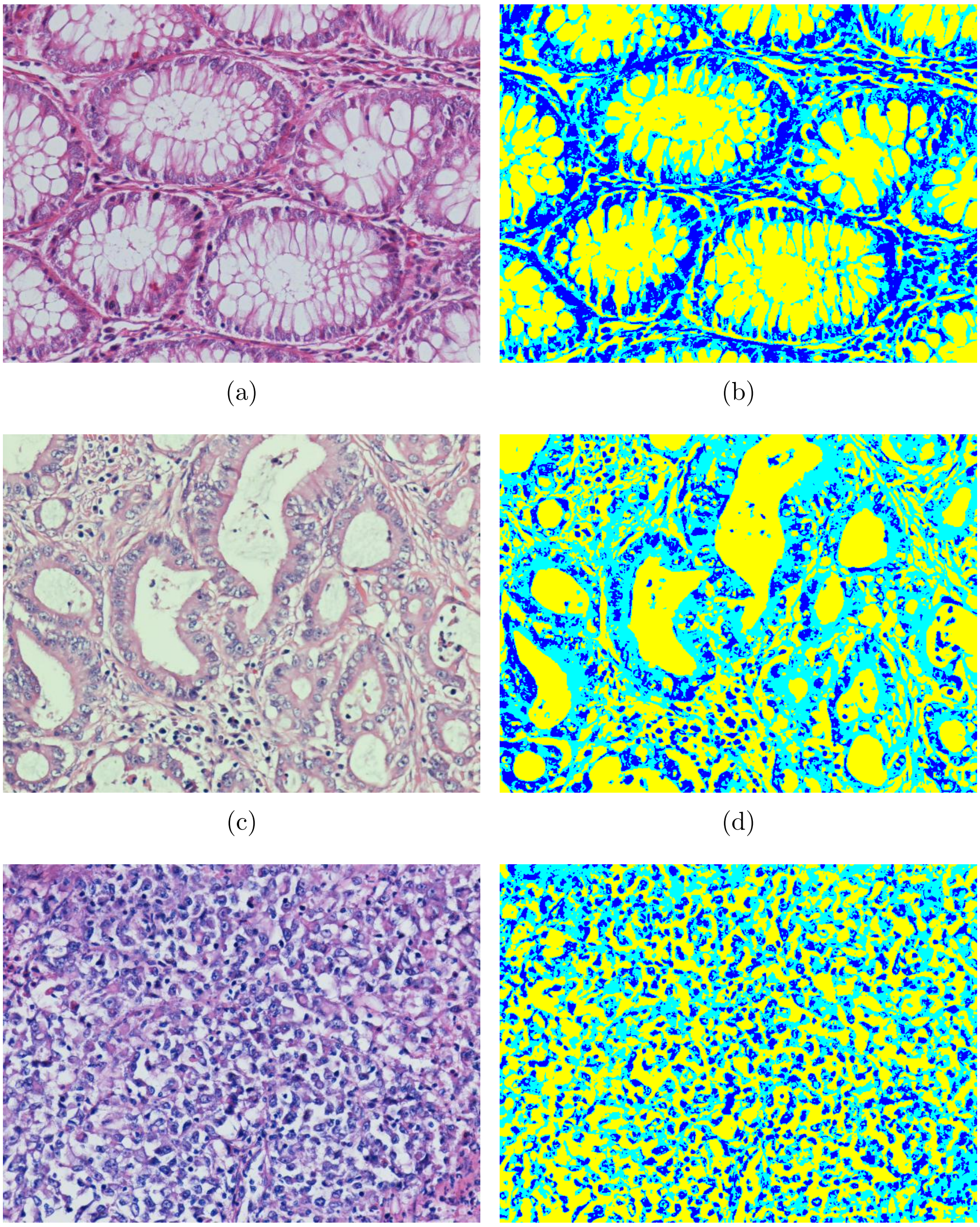
$$E = \sum_{i=1}^{k} \sum_{x_t \epsilon C_i} \|x_t - c_i\|^2 \tag{3.1}$$

The appropriate choice of k is problem and domain dependent.

We used the k-means algorithm to discriminate pixels of the nuclear, stromal, and luminal regions in a tissue image due to the fact that the H&E staining technique colors nuclei regions with dark purple, stromal regions with pink and lumen regions with white. Therefore, we select the value of k as 3. Consequently, the k-means algorithm easily separates pixels of three dissimilar regions in the image and the La*b* color conversion also increases the rate of separation. Examples of normal, low-grade cancerous, and high grade cancerous tissue images are presented in Figures 3.2(a), 3.2(c), and 3.2(e), respectively. Their corresponding clustered results are presented in Figures 3.2(b), 3.2(d), and 3.2(f), respectively. In these figures, lumen clusters are represented with yellow, stroma clusters with cyan, and nuclear clusters with blue.

3.1.2 Type assignment

After applying k-means clustering on an image, we have three disjoint regions: one for purple regions, one for pink regions, and one for white regions. In order to determine the type of clusters, the average L values of the regions are used. In the La*b* color space, L represents the light intensity of color with having 0 for black and 100 for white. Therefore, the cluster vector with the highest average L and its corresponding pixels are labeled as lumen and the darkest one and its corresponding pixels are labeled as nucleus, which typically has a purple color in the RGB space. The remaining cluster and its pixels are labeled as stroma, which has usually a pink color in the RGB space.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.2: Examples of colon tissue images : (a) normal, (c) low-grade cancerous, and (e) high-grade cancerous. Resulting k-means clusters are given in (b), (d), and (f). In this figure, lumen, stroma, and nuclear clusters are represented with yellow, cyan, and blue, respectively.

3.1.3 Salient points

At the end of the k-means algorithm, the pixels are automatically separated into three groups. Although pixel grouping provides important information about the tissue, it is hard to identify the exact boundaries of its cytological components. The reason behind this is that, due to complex nature of histopathological images, the exact boundaries of nuclear, stromal, and luminal structures are not clearly identifiable even for a human eye. Therefore, an alternative segmentation algorithm, which will be an approximation, should be used.

In our study, instead of determining the exact boundaries of tissue components, we approximately represent them by transforming each individual histological component into a circular primitive. We particularly select a circular shape for the transformation because borders of the tissue components typically are composed of curves. Moreover, circles are efficiently located on a set of pixels and they are easy to compute compared to, for example, elliptical shapes. For defining these circular objects, we make use of a technique called the *circle-fit algorithm*, implemented by our research group. This algorithm locates circles on given connected components [28, 70].

Before calling the circle-fit algorithm, a preprocessing on the pixels of each cluster is performed to decrease the effects of noise due to the incorrect assignment of pixels in the k-means clustering step. This preprocessing includes a series of morphological operations (a morphological closing followed by a morphological opening with a square structuring element of size 3) to reduce the noise in the results.

After the preprocessing step, we run the circle-fit algorithm on the white, purple, and pink clusters, separately. Circles are iteratively located on a given set of pixels that are in the same connected component. Moreover, it is possible to have some small artifacts around luminal, stromal, and nuclear regions due to incorrect quantization of pixels in the clustering step and these artifacts result in undesired circular primitives in the output. In order to reduce these artifacts, a radius threshold is employed in the circle-fit algorithm. In this study, we set the circle radius threshold to 3 for the white and pink regions and to 2 for the purple regions since nuclei are expected to be smaller than the other components. The details of the circle-fit algorithm are presented in [28, 70].

The output of the circle-fit algorithm is a set of circular primitives that approximately represent the tissue components. Figure 3.3 shows the resulting circular primitives found for the clusters given in Figure 3.2. Likewise, in this figure, yellow, cyan, and blue circles represent luminal, stromal, and nuclear components, respectively.

In our study, we use these circular primitives as *salient points*, which have a potential of carrying important biological information. A salient point is defined as a quadruple such that $S_k =< x_k, y_k, r_k, t_k >$ is the k^{th} salient point where (x_k, y_k) are the x and y coordinates of its centroid, r_k is its radius, and t_k is the type, $t_k \in \{nucleus, stroma, lumen\}$. This definition allows us to define a set of features around these salient points for quantifying and representing tissue images. Next section explains the way how we extract some intensity-based and textural features by using these salient points.

3.2 Salient Point Patterns

As explained in previous sections, we have partitioned the pixels of a tissue image into there disjoint regions and circular primitives are defined on these regions. Then, we identify these circular primitives as salient points. However, these salient points are not sufficient to be analyzed by themselves. Therefore, some quantitative features are necessary to represent salient points, and hence, to represent a tissue image. For this purpose, we propose a method to extract quantitative information around salient points to be used in cancer diagnosis and grading.

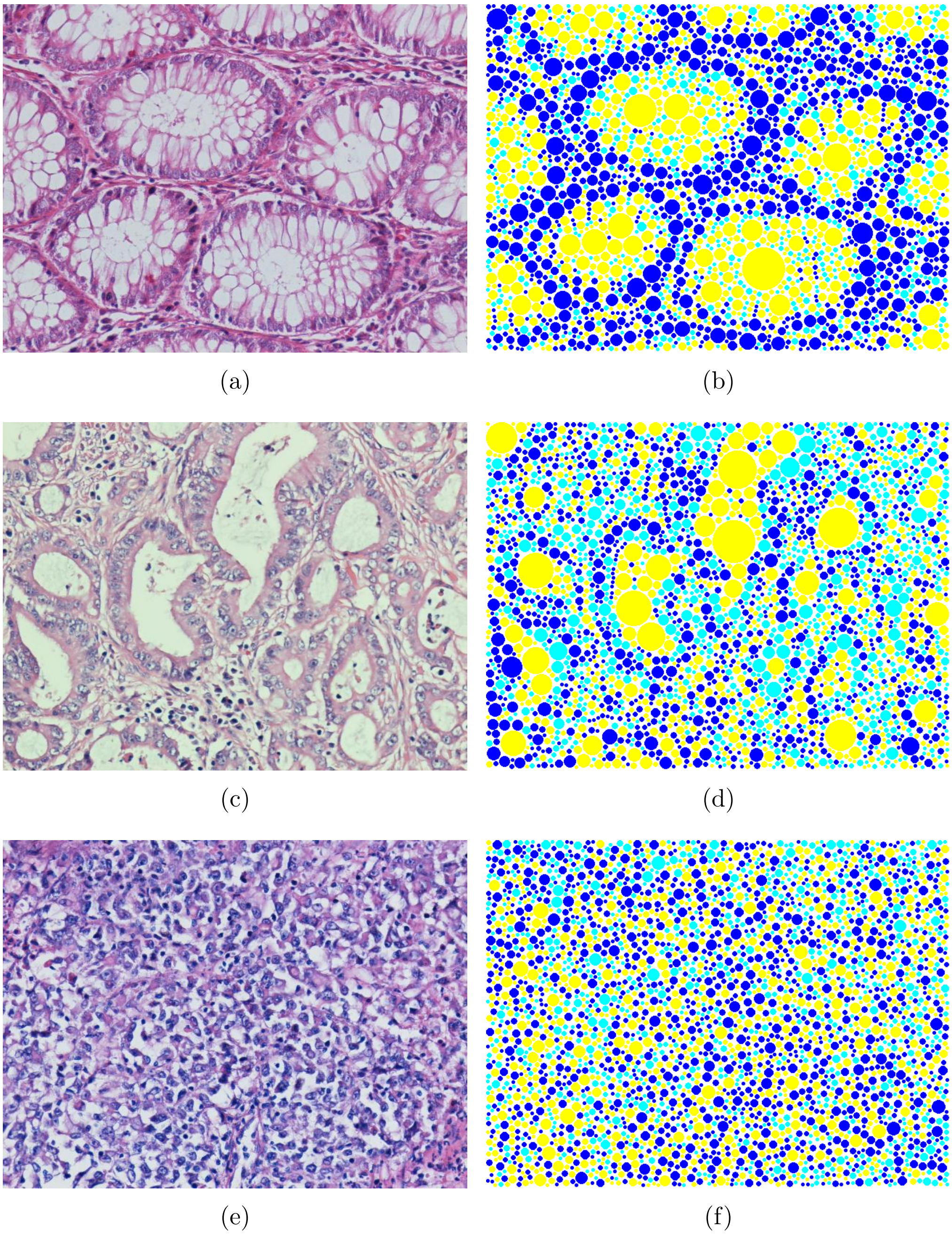

(a)
(b)
(c)
(d)
(e)
(f)

Figure 3.3: Examples of colon tissue images : (a) normal, (c) low-grade cancerous, and (e) high-grade cancerous. The output of the circle-fit algorithm are given in (b), (d), and (f). In this figure, lumen, stroma, and nuclear components are represented with yellow, cyan, and blue circles, respectively.

The previous step (Section 3.1.3) explains the definition of salient points. Let, I be the tissue image which is represented with a set of features:

$$I = \{f_k^{R'}\}_{k=1}^{K} \tag{3.2}$$

where $f_k^{R'}$ is the feature vector extracted for the k^{th} salient point (S_k) using a circular window with a radius of R'. Note that $R' = R + r_k$, where R is an external parameter selected for all of the salient points and r_k is the radius of the k^{th} salient point. K is the total number of salient points in the image. These features $f_k^{R'}$ are used to define a feature vector F. This definition will be explained towards the end of this section. We call this feature extraction pattern as Salient Point Patterns (SPP). Figure 3.4 illustrates the SPP on a sample tissue image. In this figure, the red circle, centered on the centroid of a salient point, is a circular window which is used as a mask to extract features within its area. Note that if the radius of a salient point increases, the radius of the circular window that surrounds this salient point also increases.

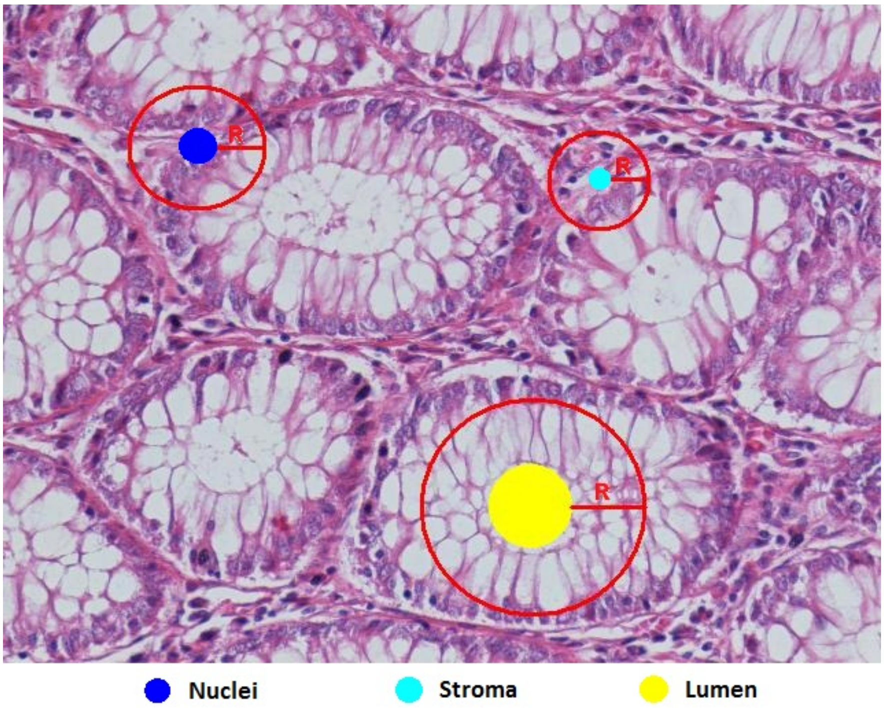


Figure 3.4: Some Salient Point Patterns (SPP) on a normal colon tissue. Here yellow, cyan, and blue circles correspond to examples of lumen, stroma, and nuclei components, respectively.

Moreover, the feature vector f, that is extracted by using a surrounding circle, could be derived by using different intensity-based or textural approaches. In this study, we employ color histogram, co-occurrence matrix, run-length matrix, local binary pattern (LBP) histogram, and Gabor filter features; these features are explained in Section 3.3. The final feature vector that represents the tissue image is constructed by accumulating each feature vector of salient points with respect to their types. Suppose that the mean and standard deviation computed from the feature vectors of t type salient points are denoted as μ_t and σ_t:

$$\mu_t = \frac{\sum_{i=1}^{n_t} f_i^{R'}}{n_t} \tag{3.3}$$

$$\sigma_t = \sqrt{\frac{\sum_{i=1}^{n_t} |f_i^{R'} - \mu_t|^2}{n_t}} \tag{3.4}$$

Here t is the salient point type such that $t \in \{nucleus, stroma, lumen\}$ and n_t is the number of salient points with type t. Note that both μ_t and σ_t are vectors. Consequently, we define a feature vector F that quantifies the entire tissue image by employing the means and standard deviations :

$$F = \{\mu_{nuclei}, \sigma_{nuclei}, \mu_{stroma}, \sigma_{stroma}, \mu_{lumen}, \sigma_{lumen}\} \tag{3.5}$$

The SPP method aims to capture the visual properties of tissue components. If we analyze Figure 3.4, it can be observed that the salient point of lumen type located in the center of glandular structure and circular window around this salient point nearly fill the glandular area. Therefore, this SPP gives information about the characteristics inside glandular structures. Moreover, an epithelial cell nucleus is usually located at the border of glandular structures. Thus, the SPP with a nucleus type helps capture characteristics around the gland boundary. Stromal structures usually correspond to connective tissue components that are not part of glandular structures. Thus, the SPP of stroma type helps characterize the regions in between the glandular structures. With three distinct types of salient points and the SPP method that provides a way to extract features around these salient points, we can employ different features. In the next section, the features that are used to represent a tissue image will be discussed.

3.3 Feature Extraction

There are many features that can be used in the SPP framework. In this study, we cover some of the intensity-based and textural features. In this section, details of the selected features are presented.

3.3.1 Color histogram features

Color histogram is a structure that models the distribution of color intensities of an image. Generally, gray level or color intensities are put into bins to construct the histogram and first order statistical features are extracted on this histogram. In this study, we employ gray level histogram to extract our intensity-based features.

For the calculation of the color histogram features, an RGB image is transformed into gray level. Then, gray intensities are quantized into N bins. It is common to observe brightness changes in tissue images. Histogram normalization is applied to reduce the effect of these brightness changes. The probability density function $h(g_i)$ of the gray level g_i, satisfying the following condition :

$$\sum_{i}^{N} h(g_i) = 1 \tag{3.6}$$

is used to describe the histogram. A feature vector is defined on the histogram by computing the mean, standard deviation, skewness, kurtosis, and entropy features, as presented in Table 3.1 [76].

Mean	$\mu \quad = \quad \sum_i^N h(g_i)g_i$
Standard deviation	$\sigma \quad = \quad \sum_i^N (g_i - \mu)^2 h(g_i)$
Skewness	$S \quad = \quad \sum_i^N (g_i - \mu)^3 h(g_i)$
Kurtosis	$K \quad = \quad \sum_i^N (g_i - \mu)^4 h(g_i)$
Entropy	$E \quad = \quad (-) \sum_i^N h(g_i) log_2 h(g_i)$

Table 3.1: The intensity-based features defined on the gray level histogram.

3.3.2 Co-occurrence matrix features

A co-occurrence matrix is used to define the second order texture measures. It considers the spatial relationship between each pair of pixels. It is initially defined by Haralick in 1973 [30]. Each of its entry specifies the number of times pixel value p_i co-occurred with pixel value p_j in a particular relationship define by a distance d and orientation θ. The co-occurrence matrix M is defined over a $w \times h$ image I, parameterized by an offset $(\triangle x,\ \triangle y)$:

$$M_{\triangle x, \triangle y}(i,j) = \sum_{p=1}^{w} \sum_{q=1}^{h} \begin{cases} 1, & \text{if } I(p,q) = i \text{ and } I(p + \triangle x, q + \triangle y) = j \\ 0, & otherwise \end{cases} \tag{3.7}$$

A pixel value of an image could be any value from 32-bit color to binary. For example, if an image is 8-bit color, as in our case, the corresponding co-occurrence matrix will be $2^8 \times 2^8$ size, which takes more memory space. Moreover, such a co-occurrence matrix is sensitive to noise in an image. Therefore, we quantize the gray intensity values into different number of bins N. A co-occurrence matrix is also rotation-variant, so the rotatin invariance is achieved by the use of a set of offsets corresponding to orientation $\theta = \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$ with the same distance d. Subsequently, for a given distance d, we accumulate each resulting co-occurrence matrix to make it invariant to rotation. The illustration of the co-occurrence matrix generation process is presented in Figure 3.5.

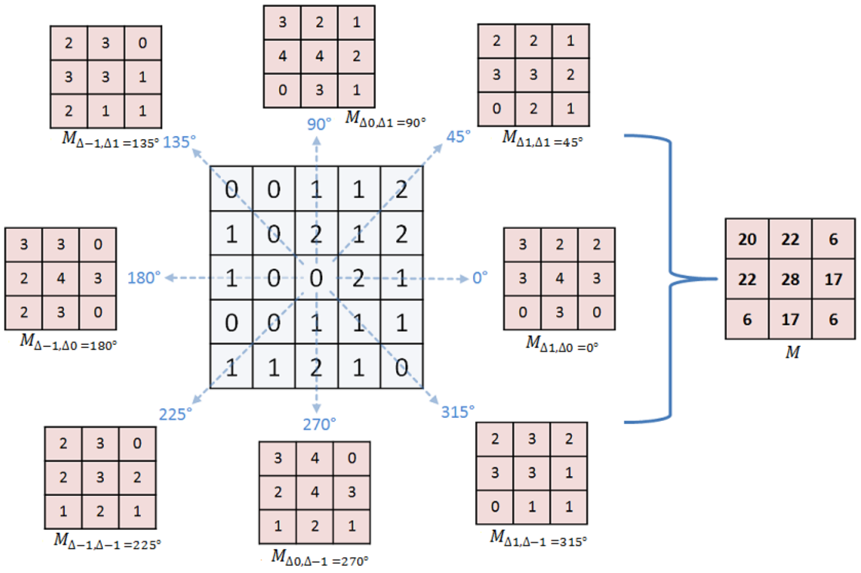

2 3 0
3 3 1
2 1 1
$M_{\Delta-1,\Delta1}$ =135°
135°
3 2 1
4 4 2
0 3 1
90°
$M_{\Delta0,\Delta1}$ =90°
2 2 1
3 3 2
0 2 1
45°
$M_{\Delta1,\Delta1}$ =45°
0 0 1 1 2
1 0 2 1 2
1 0 0 2 1
0 0 1 1 1
1 1 2 1 0
3 3 0
2 4 3
2 3 0
180°
$M_{\Delta-1,\Delta0}$ =180°
3 2 2
3 4 3
0 3 0
0°
$M_{\Delta1,\Delta0}$ =0°
20 22 6
22 28 17
6 17 6
M
225°
2 3 0
2 3 2
1 2 1
$M_{\Delta-1,\Delta-1}$ =225°
270°
3 4 0
2 4 3
1 2 1
$M_{\Delta0,\Delta-1}$ =270°
315°
2 3 2
3 3 1
0 1 1
$M_{\Delta1,\Delta-1}$ =315°

Figure 3.5: The accumulated co-occurrence matrix computed over those when d is selected as 1.

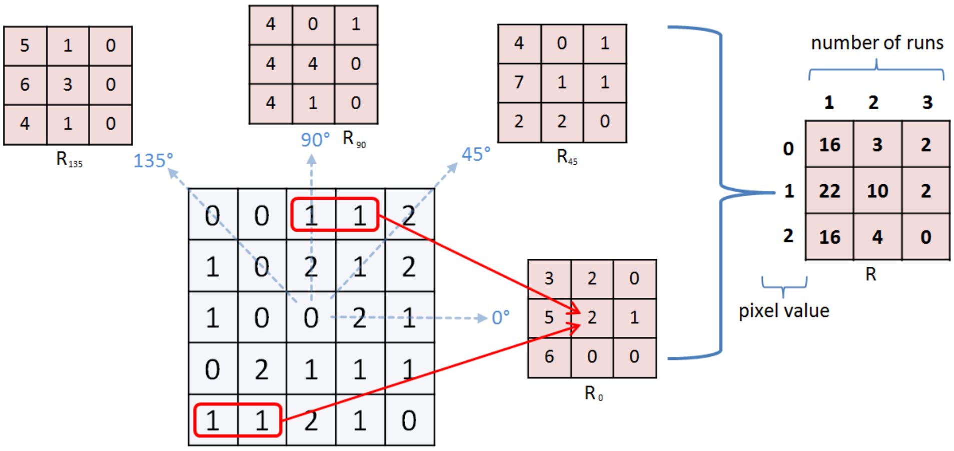
The raw co-occurrence matrix is not sufficient to describe the texture in an image. Therefore, many textural features are derived from the co-occurrence matrix. In this study, we extract six most commonly used statistical features [30] on the co-occurrence matrix. These features are *entropy* to measure randomness, *homogeneity* to characterize the image homogeneity, *correlation* to calculate the linearity of gray-level dependencies, *dissimilarity* to measure the degree of dissimilarity between pixels, *inverse difference moment* to identify the local homogeneity of the image and *maximum probability* to keep the maximum of co-occurrence matrix. Table 3.2 presents the formula of these features.

Entropy	$=$	$\sum_i \sum_j M_d(i,j) \log M_d(i,j)$
Homogeneity	$=$	$\sum_i \sum_j \frac{M_d(i,j)}{1+\lvert i-j \rvert}$
Correlation	$=$	$\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)M_d(i,j)}{\sigma_x \sigma_y}$
Dissimilarity	$=$	$\sum_i \sum_j \lvert i-j)\rvert M_d(i,j)$
Inverse difference moment	$=$	$\sum_i \sum_j \frac{M_d(i,j)}{1+\lvert i-j \rvert^2}$
Maximum probability	$=$	$max M_d(i,j)$

Table 3.2: The textural features derived from a co-occurrence matrix.

3.3.3 Run-length matrix features

Galloway proposes the use of a run-length matrix for texture representation in an image [24]. It is another way of defining higher order statistical texture features. The run-length matrix $R_\theta(i, j)$ keeps the number of runs of j-length consecutive, collinear pixels that all have the same gray value i, in the direction of θ. In this study, we compute four run-length matrices over four basic directions $\theta = \{0°, 45°, 90°, 135°\}$ and accumulate the resulting run-length matrices to define rotation-invariant matrix. Figure 3.6 illustrates the construction of the accumulated run-length matrix from a given gray level matrix. In this figure, rows represent the gray level of a run, and columns represent the length of the run.



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

R_{135}

5	1	0
6	3	0
4	1	0

R_{90}

4	0	1
4	4	0
4	1	0

R_{45}

4	0	1
7	1	1
2	2	0

R_0

3	2	0
5	2	1
6	0	0

R

	1	2	3
0	16	3	2
1	22	10	2
2	16	4	0

Figure 3.6: Run-length matrices derived from a gray level image (G = 3) in different orientations. Rows represent the gray level of a run, columns represents the length of the run. The accumulated run-length matrix is also shown.

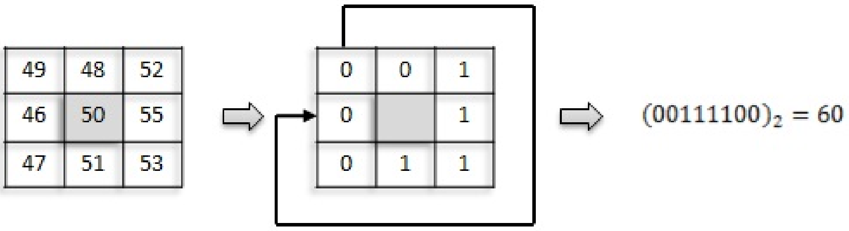
Galloway defines a set of textural features on run-length matrices. The features derived from a run-length matrix are *short run emphasis*, *long run emphasis*, *gray level nonuniformity*, *run-length nonuniformity*, and *run percentage*; those features are given in Table 3.3. Before computing the run-length matrices, image pixels are quantized into N bins to reduce the effects of noise occurred in images.

Short run emphasis	=	$\frac{\sum_i \sum_j \frac{R(i,j)}{j^2}}{n}$
Long run emphasis	=	$\frac{\sum_i \sum_j R(i,j).j^2}{n}$
Gray level nonuniformity	=	$\frac{\sum_i (\sum_j R(i,j))^2}{n}$
Run-length nonuniformity	=	$\frac{\sum_j (\sum_i R(i,j))^2}{n}$
Run percentage	=	$\frac{\sum_i \sum_j R(i,j)}{p}$

Table 3.3: The textural features derived from run-length matrices. In this table, p is the number of pixels in an image and $n = \sum_i \sum_j R(i,j)$.

3.3.4 Local binary pattern features

Local binary patterns (LBP) are a powerful method to capture local textural properties within an image [49, 50]. For a simple definition, this method compares the grayscale value of $P_{i,j}$ with those of its eight nearest neighbors $N_n (n = 1, , 8)$. The results from eight neighbors are used to form a binary number, $b_1 b_2 ... b_8$, where $b_n = 0$ if the pixel value of the n^{th} neighbor is less than that of $P_{i,j}$ and $b_n = 1$, otherwise. The computation of LBP for a given pixel is presented in Figure 3.7.



49	48	52
46	50	55
47	51	53

0	0	1
0		1
0	1	1

$(00111100)_2 = 60$

Figure 3.7: Illustration of extraction an LBP features.

The LBP operator is circular, which means that it considers the surrounding neighbors of the central pixel. Therefore, when the image is rotated, the binary pattern will only be shifted. Based on this observation, Ojala *et al.* introduce a new definition called *uniform* LBP [50]. They call an LBP uniform if it contains, at most, two bitwise 0/1 transition in its circular chain. Based on this definition, we compute the histogram of a rotation invariant LBP as illustrated in Figure 3.8. In this figure, the numbers inside circular patterns correspond to the respective bin numbers in the histogram. For example, pixel p_i having $LBP_{p_i} = 10000000_2$ and pixel p_j having $LBP_{p_j} = 01000000_2$ contribute to the same $bin = 1$ since the binary pattern of p_i is only the shifted version of p_j's binary pattern. Moreover, we add an extra bin for remaining non-uniform patterns. The construction of an LBP histogram can be effectively done with the help of a lookup table. The feature vector of length 10 is constructed using the raw LBP histogram.

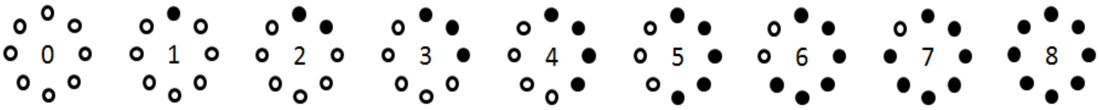


Figure 3.8: Rotation invariant binary patterns with white and black circles correspond to 0 and 1 in the output of the LBP operator. The numbers inside them correspond to the respective bin numbers.

3.3.5 Gabor filter features

Gabor filters are one of the commonly used techniques for image texture representation [40, 53, 82, 83]. Basically, Gabor filters can be described as a group of wavelets, with each wavelet capturing energy at a specific frequency and a specific direction [82]. In this study, we use the 2-D Gabor filter implementation in Matlab [37]. For a given image $I(x, y)$ with size $P \times Q$, its Gabor wavelet transform is given by a convolution:

$$G_{m,n}(x, y) = \sum_s \sum_t I(x - s, y - t)\Psi^*_{m,n}(s, t) \tag{3.8}$$

where, s and t are the filter mask size variables, m is scale, n is orientation, and $\Psi^*_{m,n}$ is a wavelet function. Since tissue images contain glandular structures in circular shapes with different sizes, there is no specific direction that glands are oriented. Therefore, we use orientation independence for better texture representation. To this end, magnitudes at each orientation for the same scale are summed up:

$$S_m(x, y) = \sum_n |G_{m,n}(x, y)| \tag{3.9}$$

To model the homogeneity of images or regions, following mean μ_m and standard deviation σ_m are computed:

$$\mu_m = \frac{\sum_x \sum_y S_m(x,y)}{P \times Q} \tag{3.10}$$

$$\sigma_m = \frac{\sum_x \sum_y (S_m(x,y) - \mu_m)^2}{P \times Q} \tag{3.11}$$

where $P \times Q$ is the total number of pixels in an image. Using μ_m and σ_m as its components, a feature vector f is derived from the Gabor filters. Four different scales and six orientations $n = \{0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6\}$ are used in the implementation and the feature vector is given by:

$$f = (\mu_0, \sigma_0, ..., \mu_4, \sigma_4) \tag{3.12}$$

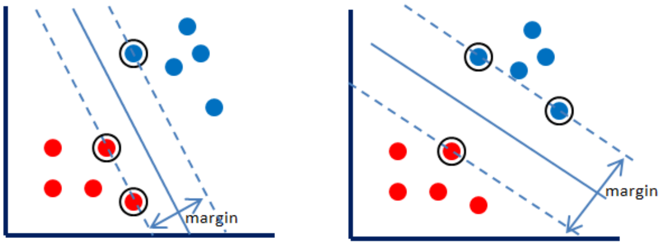
3.4 Tissue Classification

The last step of our proposed system is tissue classification. Up to now, we have built our system as to quantify a colon biopsy image with a distinctive set of features. After extracting these features, we train a classifier to classify unknown samples. In this section, we cover the classifier that is used in tissue classification and the cross-validation method for estimating the model parameters.

The success of a classification system depends on two important factors: feature definition and classifier selection. After defining a set of textural and intensity-based features on colon tissue images, we should decide a classifier to use in our system. In literature, there are many classifiers that are actively used in many medical image analysis studies such as k-nearest neighbors (KNN) [11, 36, 54, 61], decision trees [38, 65, 84], Bayesian classifiers [5, 36, 48, 61], and support vector machines (SVM) [2, 18, 73, 57, 57, 56]. We have experimented some of these classifiers in our study. Finally, we have decided to use an SVM classifier since it performs the best among the other classifiers.

3.4.1 Support vector machines (SVM)

A support vector machine (SVM) is a supervised learning technique that is used for both regression and classification based on the statistical learning theory [13]. For a simple definition, an SVM maps two sets of input data points to a high or infinite dimensional feature space and constructs a hyperplane or a set of hyperplanes, which separates these data points. New data points are then mapped to the same space and predicted to a category based on which side of the hyperplane they fall in.



(a) Smaller margin (b) Larger nargin

Figure 3.9: Two different hyperplane constructed by a support vector classifier: (a) Smaller margin and (b) larger margin.

Figures 3.9(a) and 3.9(b) present two different hyperplanes, drawn as solid lines, which separate two classes of data points. In this figure, data points closest to a hyperplane, marked with circles, are called *support vectors* and the distance between the support vectors of different classes is called *margin*. The aim of the SVM is to find an optimal hyperplane such a way that the closest member of each class are far from each other. Therefore, good separation is achieved by maximizing the margin in the SVM.

The data points given in Figures 3.9(a) and 3.9(b) are linearly separable, which means that they can be completely separable by a single line. However, there exist some data points that are not linearly separable. To handle such cases, a *kernel function* is used to transform the data points to a higher dimensional feature space in order to make it possible to separate the data. A radial basis function (RBF), a sigmoid function, and a polynomial function are some of the kernel functions that are commonly used with an SVM.

Although the kernel function is provided, it is not always possible to separate the given data points according to their categories. If the classifier constructs a model that handles every data points, this model may not be generalized well to classify unseen test samples. This may result in dramatic decrease in the classification accuracy. To overcome this problem, the SVM provides a *regularization parameter*, C, which allows user to control the trade-off between errors of the SVM on training data and margin maximization [23, 59]. Larger C values correspond to giving a higher penalty to training errors and result in constructing more strict hyperplanes. Smaller C values result in wider margins and increase training errors.

In our experiments, we used LIBSVM [10] implementation of the SVM classifier. It is available online at **http://www.csie.ntu.edu.tw/-cjlin/libsvm**.

3.4.2 Cross-validation

During the feature extraction and classification phase, there are some unknown parameters that need to be estimated. For instance, the number of bins used in the computation of gray level color histogram, the number of bins and the distance parameter in the computation of co-occurrence matrices, the radius parameter in our SPP method are some of them that we consider while extracting the features. Additionally, the selection of regularization parameter C in the SVM classifier highly affects the accuracy of the classification. For the accurate selection and generalization of these parameters, we use *k-fold cross-validation.* Cross-validation is a statistical method of evaluating and comparing the learning algorithms by partitioning the data into two segments: one used to learn the model and the other used to validate the model [35]. For the k-fold cross-validation, the training data is partitioned into k mutually exclusive subsets, which are called as the $folds$. One of these k subsets is selected as the test data, and the remaining $k-1$ are used as the training data. The classification accuracy is validated with the predefined set of parameters, such as the regularization parameter C of the SVM classifier or the number of bins N in gray level color histogram, using this particular test data, and this is repeated for each distinct subsets. The average classification accuracy is used to determine the value of parameters such that for a given predefined set of parameters, the one with the maximum cross-validation accuracy is selected. In our experiments, we use 10-fold cross-validation. Note that, after selecting the parameters, we test the learned model on a separate data set, which is not used in parameter selection at all.

Chapter 4

Experimental Results

This chapter presents the evaluation of our experiments on histopathological colon biopsies. The data set preparation process, parameter selection, the success of our proposed method, and comparisons will be explained in detail.

4.1 Experimental Setup

In our experiments, we used 3236 colon biopsy samples that are taken from 258 different patients; these patients are randomly selected and collected from the Pathology Department archives of Hacettepe School of Medicine during the years 2004-2009. The samples are composed of 5 micron-thick tissue sections that are stained with hematoxylin-and-eosin, which is the routinely used technique to stain biopsies in clinical institutions. The images of our dataset are taken with a Nikon Coolscope Digital Microscope using 20× microscope objective lens. This magnification level is high enough to obtain homogenous images and at the same time low enough to obtain images containing multiple glands and tissue components. At first, image resolution is selected as 960 × 1280. However, this size requires much more computational time. Therefore, each image is down sampled to 480 × 640 resolution, which produces both accurate classification results and relatively lower computational times.

Our dataset consists of normal, low-grade cancerous, and high-grade cancerous colon tissue images, which are examined and graded by an expert pathologist[1]. Since a support vector machine (SVM) classifier works in a supervised manner and requires training samples, we have divided our dataset into training and testing sets. The training set is used to estimate parameters and learn models. Note that the test set is not involved in any phase of training. The number of normal, low-grade cancerous, and high-grade cancerous samples in our dataset are given in Table 4.1.

	$TrainingDataset$	$TestDataset$
Normal	510	491
Low-grade cancerous	859	844
High-grade cancerous	275	257
Total	1644	1592
Patient	129	129

Table 4.1: Number of colon tissue images in the training and test sets.

4.2 Comparison Criteria

As mentioned in the previous chapter, the proposed SPP method employs local textural features to represent colon tissue images. In order to compare the effectiveness of this method, we also extract features on the entire images, grid-partition images, and around the scale invariant feature transform (SIFT) interest points. The details of this extraction are given below. In the rest of the thesis, we refer them as *EntireImageApproach*, *GridPartitionApproach*, and *SIFTPointsApproach*.

During the feature extraction process in *EntireImageApproach*, whole content of an image is used. However, in the computation of features on *GridPartitionApproach*, image is first divided into fixed size subimages and then features are extracted on each of them. Final feature vector is constructed by computing mean

[1] Prof. Dr. Cenk Sökmensuer is currently a member of Pathology Department, Hacettepe School of Medicine.

and standard deviation of the feature vectors extracted on these subimages. We consider the set of $\{10, 20, 40, 80, 160\}$ as the grid size in our experiments.

In this study, we also employ the SIFT algorithm to detect salient points within the tissue images. Since the SPP method captures textural information around salient points, SIFT salient points have a potential of representing a tissue image with its local identities. In order to analyze this, we follow the same way as in the SPP algorithm. We first detect SIFT salient points on a tissue image with the help of Matlab/C implementation of SIFT [71]. A circular window is located at each SIFT salient points and then, textural features are extracted on this circular window. Final feature vector that describes the image is constructed by taking the mean and standard deviation of the computed features. Here we use *SIFTPointsApproach* in our comparisons because of the following: The SIFT algorithm is successfully used for many types of images but it determines the salient points without using any domain specific information. On the other hand, in our salient point definition, we make use of the approximate locations of cytological tissue components, which carry domain specific information.

4.3 Results and Comparisons

4.3.1 Parameter selection

In the experiments, our aim is to classify the given tissue images with the highest possible accuracy. For this purpose, we extract features to describe the tissue images and classification is performed with these features. However, there are some unknown parameters issued in the feature extraction and classification phases. Table 4.2 presents the list of these parameters. Here, N represents the *number of bins* used in color histogram, co-occurrence matrix, and run-length matrix computations. R is the *circular radius parameter* used in the SPP computation phase. d is the distance parameter in the computation of co-occurrence matrices. w is the *grid size* parameter considered in the grid-partition images. In addition, C is the regularization or cost parameter issued in an SVM classifier. Note that we use SVM classifiers with linear kernel functions.

Moreover, 10-fold cross-validation is applied to estimate the values of unknown parameters. For this aim, the training set (1644 tissue images) is divided into 10 distinct subsets (six sets of size 164 and four sets of size 165). Each set has nearly the same number of normal, low-grade cancerous, and high grade cancerous samples. Note that the tissue images taken from the same patient are placed into the same fold. Moreover, the images in the test set is not included in any cross-validation processes.

	Parameter	Description	Values
Color histogram	N	Number of bins	{ 8, 16, 32, 64 }
Co-occurrence matrix	d	Distance	{ 3, 5, 10, 20 }
	N	Number of bins	{ 8, 16, 32, 64 }
Run-length matrix	N	Number of bins	{ 8, 16, 32, 64 }
Grid-partitioning	w	Grid size	{5, 10, 20, 40, 80, 160 }
Salient Point Pattern	R	Circular window radius	{ 10, 20, 30, 40, 50 }
SVM linear kernel	C	Regularization parameter	{ 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.0625, 0.1, 0.125, 0.25, 0.5, 1, 2, 2.5, 4, 5, 8, 10, 16, 25, 32, 50, 64, 100, 128 }

Table 4.2: The list of the parameters issued during feature extraction and classification steps.

4.3.2 Color histogram features

In this section, classification results obtained by using the gray level color histogram features will be analyzed in detail. In Table 4.3, the confusion matrix and the classification accuracies obtained by the color histogram features when the *EntireImageApproach* is used. Here, 10-fold cross-validation is performed to identify both the SVM parameter C and histogram bin number N. Maximum cross-validation accuracy (75.39%) is obtained with $C = 0.5$ and $N = 16$. The gray level color histogram features using *EntireImageApproach* give 78.83% overall classification accuracy. It can be observed that, if color histogram features, the SVM classifier distinguishes high-grade cancerous tissues better than normal and low-grade cancerous ones.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	371	101	19	75.56
	Low-grade	82	662	100	78.44
	High-grade	24	11	222	86.38
Overall accuracy					**78.83**

Table 4.3: The confusion matrix and the accuracies obtained by the color histogram features when the *EntireImageApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	451	30	10	91.85
	Low-grade	22	716	106	84.83
	High-grade	2	15	240	93.39
Overall Accuracy					**88.38**

Table 4.4: The confusion matrix and the accuracies obtained by the color histogram features when the *GridPartitionApproach* is used.

Table 4.4 presents the classification results of the color histogram features when the *GridPartitionApproach* is used. Ten-fold cross-validation selects the SVM cost parameter $C = 16$, the grid size $W = 20$, and the histogram bin number $N = 64$. Maximum cross-validation accuracy is computed as 85.58%. Compared with the *EntireImageApproach*, the overall accuracy is improved by 9.55 percent.

Table 4.5 shows the confusion matrix and classification accuracies obtained by the color histogram features that are computed using the *SIFTPointsApproach*.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	464	25	2	94.50
	Low-grade	20	719	105	85.19
	High-grade	1	19	237	92.22
Overall Accuracy					**89.20**

Table 4.5: The confusion matrix and the accuracies obtained by the color histogram features when the *SIFTPointsApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	483	6	2	98.37
	Low-grade	2	770	72	91.23
	High-grade	0	16	241	93.77
Overall Accuracy					**93.84**

Table 4.6: The confusion matrix and the accuracies obtained by the color histogram when the proposed SPP method is used.

In this case, the histogram bin $N = 64$ and the circular window radius $R = 10$ are selected. The cross-validation accuracy is measured as 88.75%. The overall accuracy is increased to 89.20%, which is slightly better than the *GridPartitionApproach*.

Table 4.6 presents the classification results obtained by using the proposed SPP method. Here, maximum cross-validation accuracy is achieved when $C = 5$, $N = 64$, and $R = 10$ are selected. Ten-fold cross-validation accuracy is measured as 91.54%. The overall accuracy is increased to 93.84%, which is the best classification accuracy compared with the aforementioned approaches. This result is statistically significant with significance level $p < 0.05$. By using the SPP method, 98.37% of normal colon tissues are classified accurately. Moreover, 91.23% of low-grade and 93.77% of high-grade cancerous tissues are correctly classified.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	447	30	14	91.04
	Low-grade	22	707	115	83.77
	High-grade	3	15	239	93.00
Overall Accuracy					**87.50**

Table 4.7: The confusion matrix and the accuracies obtained by the color histogram features when the *TypelessApproach* is used.

In order to analyze the effects of the type assignment step of the proposed SPP method, we make some experiments without considering the types of the salient points. In other words, a feature vector is computed based on all salient points regardless of their types (we will refer this method as the *TypelessApproach* thereafter). Table 4.7 presents the classification results based on this scheme. The highest cross-validation accuracy is measured as 86.86% where $C = 0.25$, $N = 64$, and $R = 5$. Here, the overall accuracy significantly decreases from 93.84% to 87.50% compared with the proposed SPP method, which alos involves the type assignment step. This shows that feature computation by considering the types of the salient points in the SPP approach provides higher classification accuracies.

4.3.3 Co-occurrence matrix features

This section presents the experimental results obtained with the co-occurrence matrix features. In Table 4.8, the confusion matrix and classification accuracies obtained by the co-occurrence matrix features are shown when the *EntireImageApproach* is used. Maximum cross-validation accuracy, 81.20%, is achieved when the number of bin $N = 32$, distance parameter $d = 10$, and the cost parameter $C = 100$. We can see that overall accuracy, 83.54%, is achieved with correctly classifying 81.47% of normal, 84.60% of low-grade cancerous, and 84.05% of high-grade cancerous samples.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	400	60	31	81.47
	Low-grade	51	714	79	84.60
	High-grade	29	12	216	84.05
Overall Accuracy					**83.54**

Table 4.8: The confusion matrix and the accuracies obtained by the co-occurrence matrix features when the *EntireImageApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	466	17	8	94.91
	Low-grade	20	724	100	85.78
	High-grade	9	32	216	84.05
Overall Accuracy					**88.32**

Table 4.9: The confusion matrix and the accuracies obtained by the co-occurrence matrix features when the *GridPartitionApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	466	20	5	94.91
	Low-grade	15	730	99	86.49
	High-grade	3	35	219	85.21
Overall Accuracy					**88.88**

Table 4.10: The confusion matrix and the accuracies obtained by the co-occurrence matrix features when the *SIFTPointsApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	470	14	7	95.72
	Low-grade	11	772	61	91.47
	High-grade	11	38	208	80.93
Overall Accuracy					**91.08**

Table 4.11: The confusion matrix and the accuracies obtained by the co-occurrence matrix features when the proposed SPP method is used.

Table 4.9 presents the results obtained with the *GridPartitionApproach*. The value for the highest cross-validation accuracy is 87.83%, achieved with the grid size $W = 40$, the cost parameter $C = 50$, the number of bin $N = 64$, and the distance $d = 10$. Compared with the *EntireImageApproach*, the overall accuracy is increased from 83.54% to 88.32%. This is consistent with our experiments that use the color histogram features.

Table 4.10 shows the confusion matrix and classification accuracies obtained by the co-occurrence matrix features that are computed by *SIFTPointsApproach*. In this case, $N = 64$, $d = 10$, and $R = 20$ are selected by considering the maximum cross-validation accuracy, 88.75%. The overall accuracy is better than *EntireImageApproach* and similar to the *GridPartitionApproach*. Since the extraction of SIFT points requires more expensive computation, one may prefer using the *GridPartitionApproach* with the co-occurrence matrix features.

Table 4.11 presents the classification results obtained by our SPP method. Ten-fold cross-validation selects $C = 16$, $N = 8$, $d = 10$, and $R = 30$. The overall accuracy is increased to 91.08% which is the highest classification accuracy obtained by using the co-occurrence matrix features. This result is statistically significant ($p < 0.05$) when it is compared with the other approaches that use the same set of features. We also examine the effects of the type assignment on classification accuracies. Table 4.12 presents the results of the *TypelessApproach*. Likewise, from the results, we observe that the type assignment to the salient points significantly increases the classification results.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	468	12	11	95.32
	Low-grade	36	715	93	84.72
	High-grade	21	30	206	80.16
Overall Accuracy					**87.25**

Table 4.12: The confusion matrix and the accuracies obtained by the co-occurrence matrix features when the *TypelessApproach* is used.

4.3.4 Run-length matrix features

In this section, classification results obtained by employing the run-length matrix features will be analyzed in detail. Table 4.13 presents the results obtained by using the run-length matrix features computed with the *EntireImageApproach*. Maximum cross-validation accuracy (72.82%) is achieved when the number of bin $N = 32$ and the cost parameter $C = 2$. The overall accuracy is 72.36% which is very low compared to the color histogram and the co-occurrence matrix features. When we analyze the confusion matrix, we observe that, the SVM classifier is unable to distinguish low-grade and high-grade cancerous tissues.

In Table 4.14, classification results for the *GridPartitionApproach* are presented. The highest cross-validation accuracy is obtained when $N = 8$, $W = 40$, and $C = 2.5$. The overall classification accuracy increases at a rate of 13.07% compared to the *EntireImageApproach* .

Table 4.15 demonstrates the results for the *SIFTPointsApproach*. The cross-validation selects $N = 8$, $R = 10$, and $C = 10$ where the corresponding cross-validation accuracy is 85.94%. It can be observed from the table that, most of the misclassified samples occur between low-grade and high-grade cancerous tissues.

With the use of the run-length matrix features computed with the SPP method, we have obtained the classification accuracies shown in Table 4.16. Maximum cross-validation accuracy is achieved when N, R, and C values are set to

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	414	39	38	84.32
	Low-grade	71	539	234	63.86
	High-grade	29	29	199	77.43
Overall Accuracy					**72.36**

Table 4.13: The confusion matrix and the accuracies obtained by the run-length matrix features when the *EntireImageApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	455	25	11	92.67
	Low-grade	56	720	68	85.31
	High-grade	50	22	185	71.98
Overall Accuracy					**85.43**

Table 4.14: The confusion matrix and the accuracies obtained by the run-length matrix features when the *GridPartitionApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	456	13	22	92.87
	Low-grade	29	706	109	83.65
	High-grade	17	32	208	80.93
Overall Accuracy					**86.06**

Table 4.15: The confusion matrix and the accuracies obtained by the run-length matrix features when the *SIFTPointsApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	478	13	0	97.35
	Low-grade	14	764	66	90.52
	High-grade	8	35	214	83.27
Overall Accuracy					**91.46**

Table 4.16: The confusion matrix and the accuracies obtained by the run-length matrix features when the proposed SPP method is used.

64, 20, and 128, respectively. The overall classification accuracy is 91.46%. We also report the results of the *TypelessApproach* (Table 4.17). Similarly, we observe a significant accuracy decrease. Note that here $N = 64$, $R = 10$, and $C = 32$ for the maximum cross-validation accuracy, 85.46%.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	460	25	6	93.69
	Low-grade	34	716	94	84.83
	High-grade	7	22	228	88.72
Overall Accuracy					**88.19**

Table 4.17: The confusion matrix and the accuracies obtained by the run-length matrix features when the *TypelessApproach* is used.

4.3.5 LBP histogram features

In this section, experimental results made for the LBP histogram features will be summarized. Table 4.18 shows the classification accuracies obtained by the LBP histogram features when the *EntireImageApproach* is used to describe tissue images. The highest cross-validation accuracy, 81.74%, is achieved when $C = 4$. From the table, we can observe that, although 92.46% of normal tissues are accurately classified, the classifier could not distinguish low-grade and high grade-cancerous tissues, leading to 82.00% overall accuracy.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	454	26	11	92.46
	Low-grade	21	652	171	77.25
	High-grade	25	29	203	78.99
Overall Accuracy					**82.22**

Table 4.18: The confusion matrix and the accuracies obtained by the LBP histogram features when the *EntireImageApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	450	17	24	91.65
	Low-grade	12	694	138	82.23
	High-grade	16	15	226	87.94
Overall Accuracy					**86.06**

Table 4.19: The confusion matrix and the accuracies obtained by the LBP histogram features when the *GridPartitionApproach* is used.

For the *GridPartitionApproach*, maximum cross-validation accuracy is 84.73% and the selected parameters are $C = 2$ and $w = 10$. Its results are given in Table 4.19. For the *SIFTPointsApproach*, the results are given in Table 4.20. The maximum cross-validation accuracy is 84.42% that selects $R = 20$ and $C = 0.05$. Both of these approaches increase the overall test accuracy to approximately 86 percent.

The results of the proposed SPP method are shown in Table 4.21. Here $R = 10$, $C = 0.125$, and the corresponding cross-validation accuracy is 89.72%. This proposed feature extraction method significantly increases the overall test accuracy to 92.53%. Similarly, when the *TypelessApproach* is used, this accuracy

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	460	12	19	93.69
	Low-grade	12	694	138	82.23
	High-grade	5	31	221	85.99
Overall Accuracy					**86.37**

Table 4.20: The confusion matrix and the accuracies obtained by the LBP histogram features when the *SIFTPointsApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	483	5	3	98.37
	Low-grade	0	754	90	89.34
	High-grade	0	21	236	91.83
Overall Accuracy					**92.53**

Table 4.21: The confusion matrix and the accuracies obtained by the LBP histogram features when the proposed SPP method is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	472	16	3	96.13
	Low-grade	9	718	117	85.07
	High-grade	6	13	238	92.61
Overall Accuracy					**89.70**

Table 4.22: The confusion matrix and the accuracies obtained by the LBP histogram features when the *TypelessApproach* is used.

decreases to 89.70%. Here the parameters are selected as $R = 10$ and $C = 0.25$ and the maximum cross-validation accuracy is 86.13%.

4.3.6 Gabor filter features

In this section, we will investigate the Gabor filter features in detail. Table 4.23 presents the results for *EntireImageApproach*. The SVM parameter C is set to 128; the maximum cross-validation accuracy is 81.51%. In this table, it is observed that the test accuracy is low. The reason behind this is that the SVM could not distinguish low-grade and high-grade cancerous tissues good enough by using the Gabor filter features. In Table 4.24, the results achieved for the *GridPartitionApproach* are presented. For this approach, when the grid size w is set to 20 and the cost parameter C is set to 50, the highest cross-validation accuracy, 86.13%, is obtained. The results for the *SIFTPointsApproach* are reported in Table 4.25. Here, the highest cross-validation accuracy, 85.52%, is obtained by when $C = 100$ and $w = 10$.

In Table 4.26, we report the results for our SPP method. The highest cross-validation accuracy is 89.29% when $R = 20$ and $C = 64$. For the Gabor filter features, the proposed method reaches the maximum overall test accuracy. When we repeat our experiments for the *TypelessApproach*, we again observe a decrease in the test set accuracies (Table 4.27). Here the parameters are selected as $R = 10$ and $C = 16$ and the cross-validation accuracy is 83.94%.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	421	51	19	85.74
	Low-grade	109	628	107	74.41
	High-grade	37	35	185	71.98
Overall Accuracy					**77.51**

Table 4.23: The confusion matrix and the accuracies obtained by the Gabor filter features when the *EntireImageApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	449	36	6	91.45
	Low-grade	58	665	121	78.79
	High-grade	14	34	209	81.32
Overall Accuracy					**83.10**

Table 4.24: The confusion matrix and the accuracies obtained by the Gabor filter features when the *GridPartitionApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	466	22	3	94.91
	Low-grade	34	714	96	84.60
	High-grade	31	53	173	67.32
Overall Accuracy					**84.99**

Table 4.25: The confusion matrix and the accuracies obtained by the Gabor filter features when the *SIFTPointsApproach* is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	471	14	6	95.93
	Low-grade	4	736	104	87.20
	High-grade	4	37	216	84.05
Overall Accuracy					**89.38**

Table 4.26: The confusion matrix and the accuracies obtained by the Gabor filter features when the proposed SPP method is used.

		Predicted			
		Normal	**Low-grade**	**High-grade**	**Accuracy**
Actual	**Normal**	461	26	4	93.89
	Low-grade	68	668	108	79.15
	High-grade	25	46	186	72.37
Overall Accuracy					**82.60**

Table 4.27: The confusion matrix and the accuracies obtained by the Gabor filter features when the *TypelessApproach* is used.

4.3.7 Parameter analysis

In this section, we will analyze the effects of the our model parameters to the classification accuracy. In our experiments, we analyze this effect for each parameter, fixing the remaining ones to the previously selected values. Besides the SVM parameter C, the color histogram features have the number of bins N and the circular window radius R. In Figures 4.1 and 4.2, the accuracies as a function of these parameters are shown. Here we observe that the circular window radius R has a larger effect on the accuracies.

For the co-occurrence matrix features, our SPP method involves there parameters: The number of bins N, the distance d, and the circular window radius R. The accuracies obtained as a function of these parameters are given in Figures 4.3, 4.4, and 4.5, respectively. Here we observe that the parameters N slightly affects the results whereas the other two have larger affects on the accuracies.

Co-occurrence features produces the best classification accuracy when SPP method is used during the feature extraction phase. Classification results indicate that the maximum cross-validation accuracy is achieved when $N = 8$, $d = 10$, and $R = 30$. In order to analyze the optimal values of these parameters, we first fixed the distance parameter d and R. Then, we analyzed the cross validation accuracies by varying the number of bin parameter N. Figure 4.3 presents the results obtained by using co-occurrence features with SPP method. From the table, we can derive that maximum cross-validation accuracy is achieved when

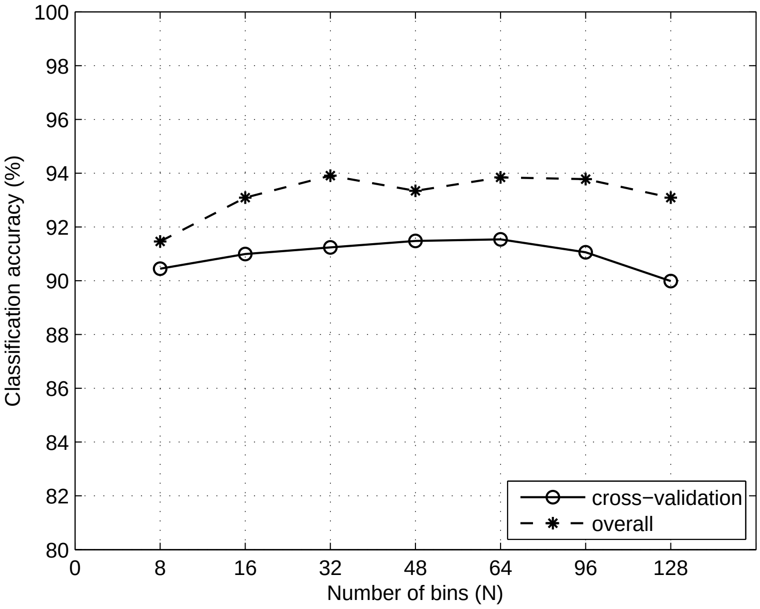


Figure 4.1: Classification accuracies as a function of the number of bins N when the SPP method uses the color histogram features. Here the circular window radius R is fixed to 10.

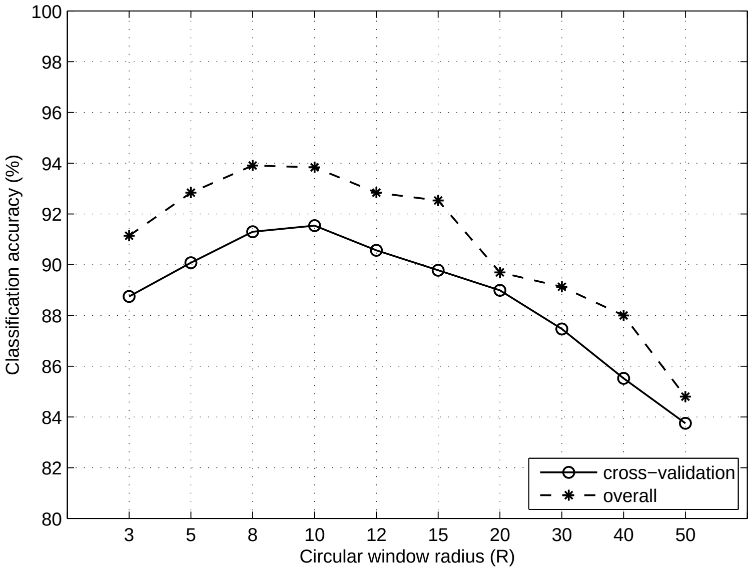


Figure 4.2: Classification accuracies as a function of the circular window radius R when the SPP method uses the color histogram features. Here the number of bins N is fixed to 64.

$N = 8$. Therefore, we fixed $N = 8$. Then, we performed analysis on the distance parameter d. Figure 4.4 presents the cross-validation accuracies obtained by varying the distance parameter. Here, cross-validation accuracy is peaked at $d = 10$. Hence, we set the distance parameter d to 10. Finally, we analyze the optimal value of the circular radius parameter R. Figure 4.5 shows the result obtained by varying the circular radius parameter R. Here, cross-validation accuracy reaches the highest value when $R = 30$. To sum up, if co-occurrence features extracted by SPP are used to classify the tissue images, SVM performs best when $N = 8$, $d = 10$, and $R = 30$.

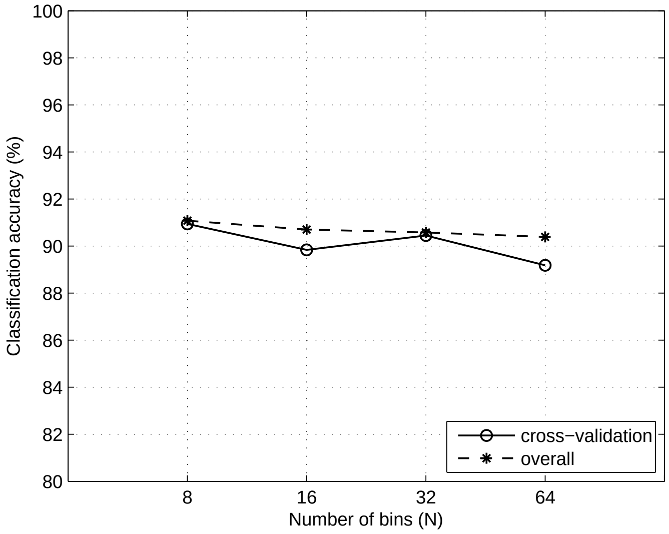


Figure 4.3: Classification accuracies as a function of the number of bins N when the SPP method uses the co-occurrence matrix features. Here the distance parameter d and the circular window radius parameter R are fixed to 10 and 30, respectively.

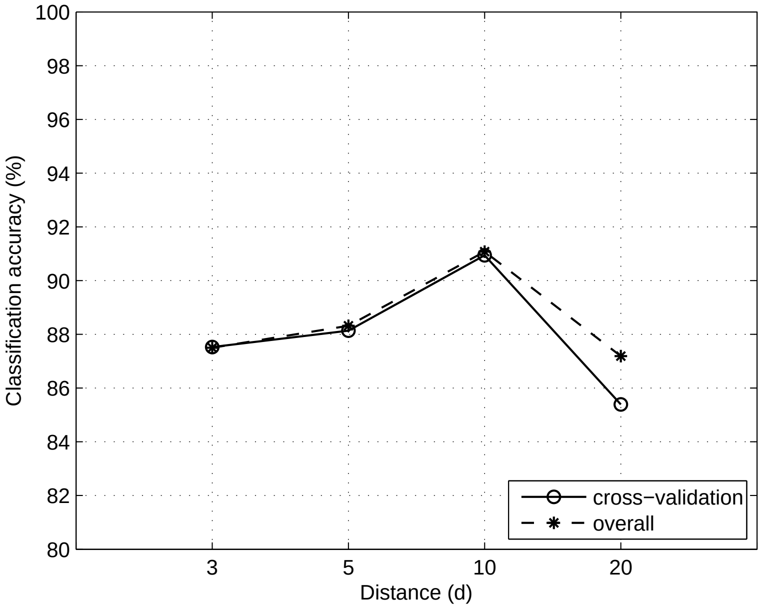


Figure 4.4: Classification accuracies as a function of the distance d when the SPP method uses the co-occurrence matrix features. Here the number of bins parameter N and the circular window radius parameter R are fixed to 8 and 30, respectively.

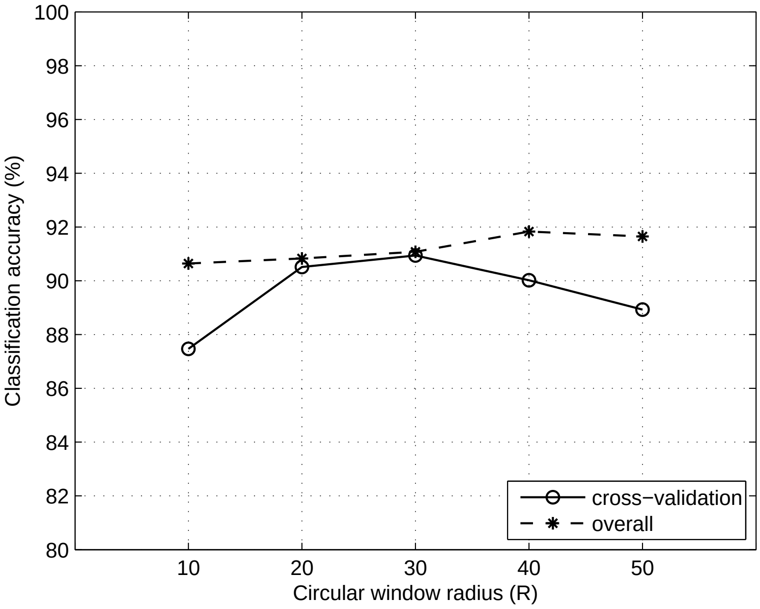


Figure 4.5: Classification accuracies as a function of the circular window radius R when the SPP method uses the co-occurrence matrix features. Here the number of bins parameter N and the distance parameter d are fixed to 8 and 10, respectively.

The SPP method has two parameters for the run-length matrix features. These are the number of bin parameter N and the circular window radius parameter. The analysis of these parameters are given in Figure 4.6 and 4.7. In these figures, we observe that the selection of the number of bins is important for obtaining higher accuracies.

For the LBP histogram and Gabor filter features, the SPP method uses only one parameter, which is the circular window radius R. Figure 4.8 and 4.9 show the accuracies as a function of R for the LBP histogram and Gabor filter features, respectively. We observe that for both of these features, the R value do not too much change the accuracy.

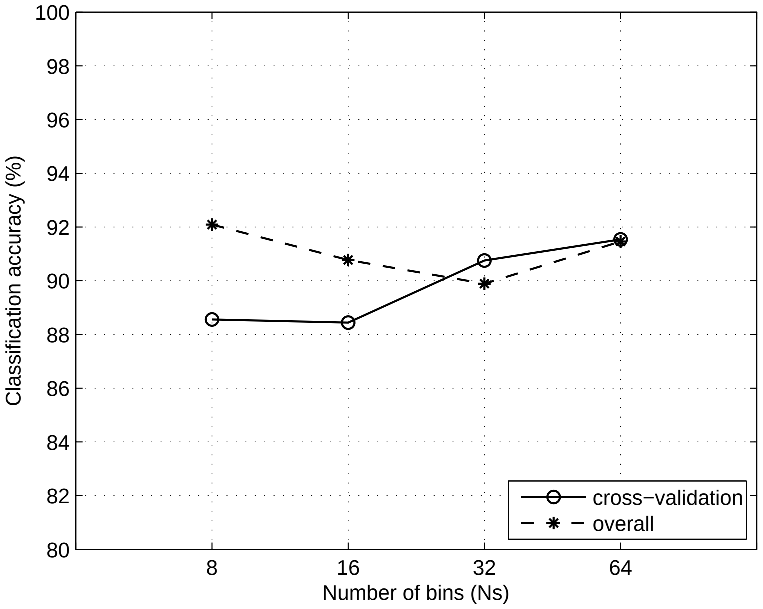


Figure 4.6: Classification accuracies as a function of the number of bins N when the SPP method uses the run-length matrix features. Here the circular window radius R is fixed 20.

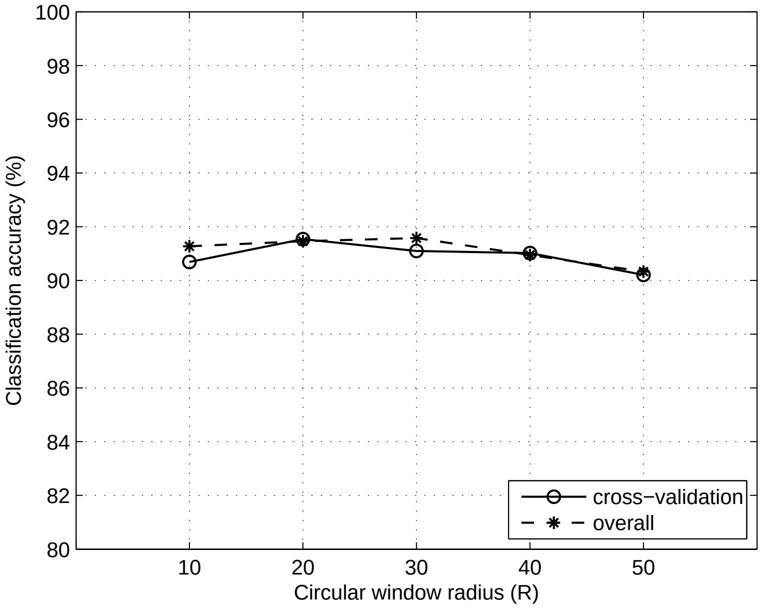


Figure 4.7: Classification accuracies as a function of the circular window radius R when the SPP method uses the run-length matrix features. Here the number of bins N is fixed 64.

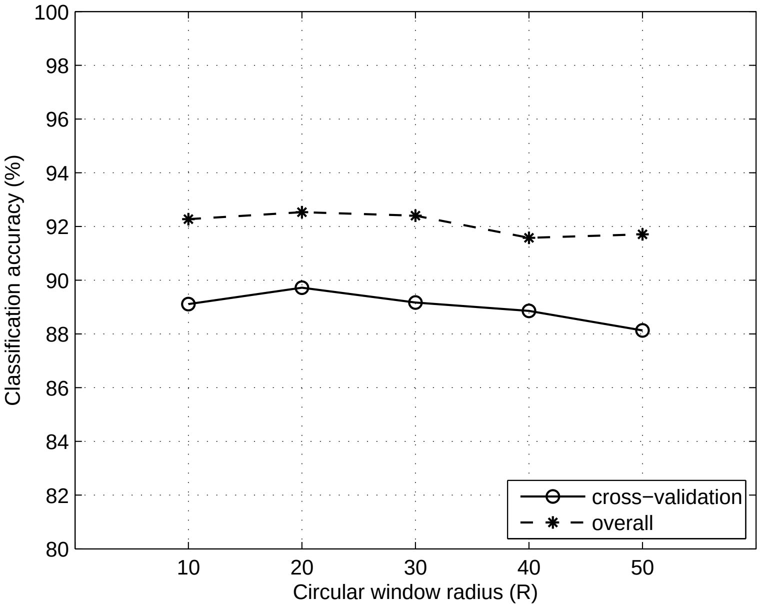


Figure 4.8: Classification accuracies as a function of the circular window radius R when the SPP method uses the LBP histogram features.

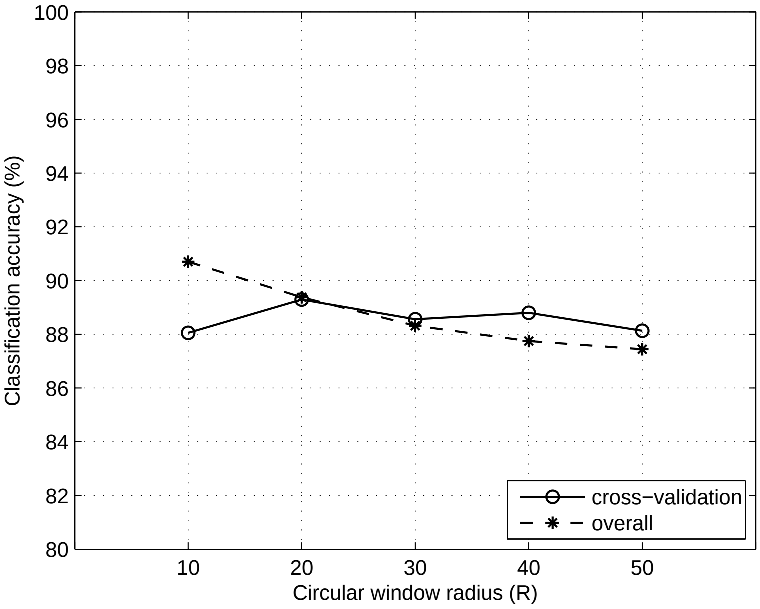


Figure 4.9: Classification accuracies as a function of the circular window radius R when the SPP method uses the Gabor filter features.

4.4 Discussion

In this section, we will discuss the experimental results.

4.4.1 Parameters

The parameter selection is one of the most important factors that highly affects the classification results. For the k-means step, we have selected k as 3 since the hematoxylin-and-eosin staining leads to three main colors in tissues, which are purple, pink, and white. Therefore, we implicitly choose this parameter value. For the salient point definition step, the minimum radius threshold used by the circle-fit algorithm is set to 2 for the nuclei and 3 for the lumen and stroma types. This selection is based on our observations and experimental results. We see that these threshold parameters are selected large enough to eliminate noise arising from the color quantization step and yield representative circular primitives.

We employed 10-fold cross-validation for the remaining parameters that appear either in the feature extraction step or classification step. Note that folds are not randomly determined since there exist many tissue images belonging to the same patient. Therefore, we determine these folds considering the patients. In Section 4.3.7, we present the list of parameters and their values selected by the 10-fold cross-validation.

4.4.2 Features

In our experiments, we have analyzed five different textural features to represent colon tissue images. For comparison, we have used four different approaches: (1) The *EntireImageApproach*, which uses an entire image to define their features, (2) the *GridPartitionApproach* which divides an image into grids, extracts features on these grids and then aggregates the features of the grids to obtain the feature vector of the image, (3) the *SIFTPointsApproach*, which finds the salient points on an image, extracts features around these salient points, and then aggregates the features, and (4) the *TypelessApproach* that follows the same steps with the proposed method except the type assignment step. The feature-based comparisons are summarized in Tables 4.29-4.32.

In these tables, we observe that the lowest accuracy is obtained when the *EntireImageApproach* is used. The results of the other three approaches (*GridPartitionApproach*, *SIFTPointsApproach*, and *TypelessApproach*) are usually more or less the same. For the LBP histogram features, the *TypelessApproach* gives higher accuracies to the other two. The proposed SPP method significantly improves the results of these algorithms for all of the features. This indicates the effectiveness of making use of biologically meaningful salient points in feature extraction. These tables (as well as the confusion matrices given in Tables 4.3-4.27) show that all these algorithms confuse low-grade and high-grade cancerous tissues the most. This decreases the overall accuracies.

	Normal	**Low-grade**	**High-grade**	**Overall**
Our SPP method	98.37	91.23	93.77	93.84
EntireImageApproach	75.56	78.44	86.38	78.83
GridPartitionApproach	91.85	84.83	93.39	88.38
SIFTPointsApproach	94.50	85.19	92.22	89.20
TypelessApproach	91.04	83.77	93.00	87.50

Table 4.28: Classification results obtained for the color histogram features.

	Normal	**Low-grade**	**High-grade**	**Overall**
Our SPP method	95.72	91.47	80.93	91.08
EntireImageApproach	81.47	84.60	84.05	83.54
GridPartitionApproach	94.91	85.78	84.05	88.32
SIFTPointsApproach	94.91	86.49	85.21	88.88
TypelessApproach	95.32	84.72	80.16	87.25

Table 4.29: Classification results obtained for the co-occurrence matrix features.

	Normal	**Low-grade**	**High-grade**	**Overall**
Our SPP method	97.35	90.52	83.27	91.46
EntireImageApproach	84.32	63.86	77.43	72.36
GridPartitionApproach	92.67	85.31	71.98	85.43
SIFTPointsApproach	92.87	83.65	80.93	86.06
TypelessApproach	93.69	84.83	88.72	88.19

Table 4.30: Classification results obtained for the run-length matrix features.

	Normal	**Low-grade**	**High-grade**	**Overall**
Our SPP method	98.37	89.34	91.83	92.53
EntireImageApproach	92.46	77.25	78.99	82.22
GridPartitionApproach	91.65	82.23	87.94	86.06
SIFTPointsApproach	93.69	82.23	85.99	86.37
TypelessApproach	96.13	85.07	92.61	89.70

Table 4.31: Classification results obtained for the LBP histogram features.

	Normal	**Low-grade**	**High-grade**	**Overall**
Our SPP method	95.93	87.20	84.05	89.38
EntireImageApproach	85.74	74.41	71.98	77.51
GridPartitionApproach	91.45	78.79	81.32	83.10
SIFTPointsApproach	94.91	84.60	67.32	84.99
TypelessApproach	93.89	79.15	72.37	82.60

Table 4.32: Classification results obtained for the Gabor filter features.

Chapter 5

Conclusion and Future Work

Computer aided diagnosis (CAD) systems have a potential to offer more stable and objective framework to pathologists for helping their decision making. Many studies have been proposed to develop such CAD systems for automated cancer diagnosis and grading, especially based on textural or structural tissue image analysis. Although these approaches provide promising results for different types of tissues, they are still incapable of using potential biological information carried by the tissue components. However, these tissue components can help better quantify the tissue images.

In this thesis, we proposed a new textural method, called Salient Point Patterns (SPP), for the utilization of tissue components to represent histopathological images of colon tissues. In the first step of this method, tissue images are transformed to the La*b* color space and their pixels are quantized into three disjoint groups by the k-means clustering algorithm. These groups correspond to nuclei (purple regions), stroma (pink regions), and lumina (white regions). In its next step, circular primitives are separately defined on these regions using the circle-fit algorithm. We call these circular primitives as *salient points*, each of which has a type (nucleus, stroma, or lumen), a radius, and a location. Afterwards, a circular window centered at the centroid of a salient point is used as a mask to extract textural or intensity based features within the window area. This feature extraction framework is called as Salient Point Patterns (SPP).

Our proposed method can be used for different feature types. In our experiments, we analyzed five different intensity-based and textural features including the color histogram, co-occurrence matrix, run-length matrix, local binary patterns (LBP) histogram and Gabor filters features. A support vector machine (SVM) with a linear kernel is used to classify tissue images into normal, low-grade cancerous, and high-grade cancerous classes. Ten-fold cross-validation is applied on training samples to estimate the parameters associated with the feature extraction and classification steps. For these features, the results obtained by the proposed method are summarized in Table 5.1.

	Normal	Low-grade	High-grade	Overall Accuracy
Color histogram	98.37	91.23	93.77	93.84
LBP histogram	98.37	89.34	91.83	92.53
Co-occurrence matrix	95.72	91.47	80.93	91.08
Run-length matrix	97.35	90.52	83.27	91.46
Gabor filters	95.93	87.20	84.05	89.38

Table 5.1: Classification results obtained by the proposed Salient Point Patterns (SPP) method for different features.

Experimental results show that the proposed SPP method improves the classification results of those obtained by the *EntireImageApproach*, *GridPartitionApproach*, and *SIFTPointsApproach*, indicating the effectiveness of defining features on the biologically meaningful salient points.

The main contribution of this thesis is the following: it offers a new feature extraction scheme that make use of the tissue components, which carry biologically important information, to describe histopathological images for cancer diagnosis and grading. The use of domain specific knowledge and mapping this knowledge to the computer environment provide us to develop more reliable and stable CAD systems.

As future work, different sets of textural features would be employed and analyzed with the SPP method. Additionally, the detailed parameter analysis can be made to improve the classification accuracy. Moreover, SPP method can be applied on other types of cancer, including prostate cancer and skin cancer.

Bibliography

[1] What is cancer? www.cancer.org/acs/groups/cid/documents/webcontent/003111-pdf.pdf, 2010. Accessed at : 17/07/2011.

[2] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir. Color Graphs for Automated Cancer Diagnosis and Grading. *IEEE Transactions on Biomedical Engineering*, 57(3):665–674, 2010.

[3] A. Andrion, C. Magnani, P. G. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, and B. Terracini. Malignant mesothelioma of the pleura: interobserver variability. *Journal of Clinical Pathology*, 48(9):856–860, 1995.

[4] G. Antoch, F. M. Vogt, L. S. Freudenberg, F. Nazaradeh, S. C. Goehde, J. Barkhausen, G. Dahmen, A. Bockisch, J. F. Debatin, and S. G. Ruehm. Whole-body dual-modality pet/ct and whole-body mri for tumor staging in oncology. *JAMA: The Journal of the American Medical Association*, 290(24):3199–3206, 2003.

[5] P. A. Bromiley, N. A. Thacker, M. L. J. Scott, M. Pokric, A. J. Lacey, and T. F. Cootes. Bayesian and non-bayesian probabilistic models for medical image analysis. *Image and Vision Computing*, 21(10):851 – 864, 2003.

[6] M. C., D. Fat, and J. Boerma. The global burden of disease: 2004 update. geneva: World health organization, 2006.

[7] J. Caicedo, A. Cruz, and F. Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Artificial Intelligence in Medicine*,

volume 5651 of *Lecture Notes in Computer Science*, pages 126–135. Springer Berlin / Heidelberg, 2009.

[8] J. Caicedo, F. Gonzalez, and E. Romero. A semantic content-based retrieval method for histopathology images. In *Information Retrieval Technology*, volume 4993 of *Lecture Notes in Computer Science*, pages 51–60. Springer Berlin / Heidelberg, 2008.

[9] W. Chan and K. H. Fu. Value of routine histopathological examination of appendices in hong kong. *Journal of Clinical Pathology*, 40(4):429–433, 1987.

[10] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[11] C. Christodoulou, C. Pattichis, M. Pantziaris, and A. Nicolaides. Texture-based classification of atherosclerotic carotid plaques. *IEEE Transactions on Medical Imaging*, 22(7):902 –912, 2003.

[12] I. S. Cook and C. E. Fuller. Does histopathological examination of breast reduction specimens affect patient management and clinical follow up? *Journal of Clinical Pathology*, 57(3):286–289, 2004.

[13] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[14] G. Daz and E. Romero. Histopathological image classification using stain component features on a plsa model. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 6419, pages 55–62. Springer Berlin / Heidelberg, 2010.

[15] C. Demir, S. Gultekin, and B. Yener. Learning the topological properties of brain tumors. *IEEE-ACM Transactions On Computational Biology And Bioinformatiocs*, 2(3):262–270, 2005.

[16] J. Diamond, N. H. Anderson, P. H. Bartels, R. Montironi, and P. W. Hamilton. The use of morphological characteristics and texture analysis in the

identification of tissue composition in prostatic neoplasia. *Human Pathology*, 35(9):1121 – 1131, 2004.

[17] E. Domingo, P. Laiho, M. Ollikainen, M. Pinto, L. Wang, A. J. French, J. Westra, T. Frebourg, E. Espn, M. Armengol, R. Hamelin, H. Yamamoto, R. M. W. Hofstra, R. Seruca, A. Lindblom, P. Peltomki, S. N. Thibodeau, L. A. Aaltonen, and S. Schwartz. Braf screening as a low-cost effective strategy for simplifying hnpcc genetic testing. *Journal of Medical Genetics*, 41(9):664–668, 2004.

[18] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro.*, pages 1284–1287, 2007.

[19] A. Esgiar, R. Naguib, B. Sharif, M. 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*, 2(3):197 –203, 1998.

[20] A. Esgiar, R. Naguib, B. Sharif, M. Bennett, and A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):54 –58, 2002.

[21] L. Fielding, P. Arsenault, P. Chapuis, O. Dent, B. Gathright, J. Hardcastle, P. Hermanek, J. Jass, and R. Newland. Clinicopathological staging for colorectal cancer: An international documentation system (IDS) and an international comprehensive anatomical terminology (ICAT). *Journal Of Gastroenterology And Hepatology*, 6(4):325–344, 1991.

[22] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harb Protoc*, 2008(5), 2008.

[23] F. Friedrichs and C. Igel. Evolutionary tuning of multiple svm parameters. *Neurocomputing*, 64:107 – 117, 2005.

[24] M. M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2):172 – 179, 1975.

[25] J. Gil, H. Wu, and B. Y. Wang. Image analysis and morphometry in the diagnosis of breast cancer. *Microscopy Research and Technique*, 59(2):109–118, 2002.

[26] F. Gress, K. Gottlieb, S. Sherman, and G. Lehman. Endoscopic ultrasonographyguided fine-needle aspiration biopsy of suspected pancreatic cancer. *Annals of Internal Medicine*, 134(6):459–464, 2001.

[27] C. Gunduz, B. Yener, and S. H. Gultekin. The cell graphs of cancer. *Bioinformatics*, 20(1):i145–i151, 2004.

[28] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer. Automatic segmentation of colon glands using object-graphs. *Medical Image Analysis*, 14(1):1–12, 2010.

[29] P. W. Hamilton, P. H. Bartels, D. Thompson, N. H. Anderson, R. Montironi, and J. M. Sloan. Automated location of dysplastic fields in colorectal histology using image texture analysis. *The Journal of Pathology*, 182(1):68–75, 1997.

[30] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610 –621, 1973.

[31] C. Heyn, J. A. Ronald, S. S. Ramadan, J. A. Snir, A. M. Barry, L. T. MacKenzie, D. J. Mikulis, D. Palmieri, J. L. Bronder, P. S. Steeg, T. Yoneda, I. C. MacDonald, A. F. Chambers, B. K. Rutt, and P. J. Foster. In vivo mri of cancer cell fate at the single-cell level in a mouse model of breast cancer metastasis to the brain. *Magnetic Resonance in Medicine*, 56(5):1001–1010, 2006.

[32] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50(6):697 –704, 2003.

[33] P. Jensen, M. R. Krogsgaard, J. Christiansen, O. Brndstrup, A. Johansen, and J. Olsen. Observer variability in the assessment of type and dysplasia of colorectal adenomas, analyzed using kappa statistics. *Diseases of the Colon Rectum*, 38:195–198, 1995.

[34] D. Keysers, J. Dahmen, H. Ney, B. Wein, and T. Lehmann. Statistical framework for model-based image retrieval in medical applications. *Journal of Electronic Imaging*, 12(1):59–68, 2003.

[35] R. Koavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artifical Intelligence (IJCAI)*, pages 6687 –6690, 1995.

[36] J. Kong, O. Sertel, H. Shimada, K. Boyer, J. Saltz, and M. Gurcan. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. *Pattern Recognition*, 42(6):1080 – 1092, 2009.

[37] P. Kovesi. Code for convolving an image with a bank of log-gabor filters. http://www.csse.uwa.edu.au/ pk/research /matlabfns/PhaseCongruency/gaborconvolve.m.

[38] W.-J. Kuo, R.-F. Chang, D.-R. Chen, and C. C. Lee. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Research and Treatment*, 66:51–57, 2001.

[39] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[40] W. Y. Ma and B. S. Manjunath. Texture features and learning similarity. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:425, 1996.

[41] K. Masood and N. Rajpoot. Texture based classification of hyperspectral colon biopsy samples using clbp. In *IEEE International Symposium on Biomedical Imaging: from Nano to Macro*, pages 1011 –1014, 2009.

[42] K. Masood, N. Rajpoot, K. Rajpoot, and H. Qureshi. Hyperspectral colon tissue classification using morphological analysis. In *International Conference on Emerging Technologies(ICET)*, pages 735 –741, 2006.

[43] H. Mller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1 – 23, 2004.

[44] E. C. M. Mommers, N. Poulin, J. Sangulin, C. J. L. M. Meijer, J. P. A. Baak, and P. J. van Diest. Nuclear cytometric changes in breast carcinogenesis. *The Journal of Pathology*, 193(1):33–39, 2001.

[45] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. *Medical Image Analysis*, 14(4):617 – 629, 2010.

[46] R. Montironi, R. Mazzuccheli, M. Scarpelli, A. Lopez-Beltran, G. Fellegara, and F. Algaba. Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. *BJU International*, 95(8):1146–1152, 2005.

[47] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information. In *MIAAB Workshop*, 2007.

[48] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287, 2008.

[49] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996.

[50] T. Ojala, M. Pietikinen, and T. Menp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.

[51] G. Paschos. Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE Transactions on Image Processing*, 10(6):932 –937, 2001.

[52] B. Ponder. Genetic testing for cancer risk. *Science*, 278(5340):1050–1054, 1997.

[53] R. Porter and N. Canagarajah. Robust rotation-invariant texture classification: wavelet, gabor filter and gmrf based schemes. *Vision, Image and Signal Processing*, 144(3):180 –188, 1997.

[54] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2008*, volume 5242 of *Lecture Notes in Computer Science*, pages 196–204. 2008.

[55] K. M. Rajpoot, N. M. Rajpoot, and M. J. Turner. Hyperspectral colon tissue cell classification. 2004.

[56] S. Raza, R. Parry, Y. Sharma, Q. Chaudry, R. Moffitt, A. Young, and M. Wang. Automated classification of renal cell carcinoma subtypes using bag-of-features. In *IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBC)*, pages 6749 –6752, 2010.

[57] S. Raza, Y. Sharma, Q. Chaudry, A. Young, and M. Wang. Automated classification of renal cell carcinoma subtypes using scale invariant feature transform. In *IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBC)*, pages 6687 –6690, 2009.

[58] R. Rubin, D. Strayer, E. Rubin, and J. McDonald. *Rubin's Pathology: Clinicopathological Foundation of Medicine.* Lippincott William and Wilkins, 2007.

[59] M. Rychetsky, S. Ortmann, and M. Glesner. Support vector approaches for engine knock detection. In *International Joint Conference on Neural Networks (IJCNN)*, volume 2, pages 969 –974 vol.2, 1999.

[60] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, J. Saltz, and M. Gurcan. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. *Journal of Signal Processing Systems*, 55:169–183, 2009.

[61] O. Sertel, J. Kong, H. Shimada, U. Catalyurek, J. Saltz, and M. Gurcan. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognition*, 42(6):1093 – 1103, 2009.

[62] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick. Assert: A physician-in-the-loop content-based retrieval system for hrct image databases. *Computer Vision and Image Understanding*, 75(1-2):111 – 132, 1999.

[63] R. Siegel, E. Ward, O. Brawley, and A. Jemal. Cancer statistics, 2011. *CA: A Cancer Journal for Clinicians*, 61(4):212–236, 2011.

[64] R. A. Smith, V. Cokkinides, A. C. von Eschenbach, B. Levin, C. Cohen, C. D. Runowicz, S. Sener, D. Saslow, and H. J. Eyre. American cancer society guidelines for the early detection of cancer. *CA: A Cancer Journal for Clinicians*, 52(1):8–22, 2002.

[65] A. Sousa, M. Dinis-Ribeiro, M. Areia, and M. Coimbra. Identifying cancer regions in vital-stained magnification endoscopy images using adapted color histograms. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 681 –684, 2009.

[66] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. 1905(1):861–870, 1993.

[67] 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*, 26(10):1366 –1378, 2007.

[68] H. Tang, R. Hanka, and H. Ip. Histological image retrieval based on semantic content analysis. *IEEE Transactions on Information Technology in Biomedicine*, 7(1):26 –36, 2003.

[69] A. Tosun and C. Gunduz-Demir. Graph run-length matrices for histopathological image segmentation. *IEEE Transactions on Medical Imaging*, 30(3):721 –732, 2011.

[70] A. B. 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*, 42(6):1104 – 1112, 2009.

[71] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. http://www.vlfeat.org/.

[72] S. Waheed, R. Moffitt, Q. Chaudry, A. Young, and M. Wang. Computer aided histopathological classification of cancer subtypes. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 503 –508, 2007.

[73] Y. Wang, D. Crookes, O. Eldin, S. Wang, P. Hamilton, and J. Diamond. Assisted diagnosis of cervical intraepithelial neoplasia (cin). *IEEE Journal of Selected Topics in Signal Processing*, 3(1):112 –121, 2009.

[74] B. Weyn, G. Van De Wouwer, M. Koprowski, A. Van Daele, K. Dhaene, P. Scheunders, W. Jacob, and E. Van Marck. Value of morphometry, texture analysis, densitometry, and histometry in the differential diagnosis and prognosis of malignant mesothelioma. *The Journal of Pathology*, 189(4):581–589, 1999.

[75] B. Weyn, G. van de Wouwer, S. Kumar-Singh, A. van Daele, P. Scheunders, E. van Marck, and W. Jacob. Computer-assisted differential diagnosis of

malignant mesothelioma based on syntactic structure analysis. *Cytometry*, 35(1):23–29, 1999.

[76] M. Wiltgen, A. Gerger, and S. J. Tissue contour analysis of bening. *International Journal of Medical Informatics*, 69:17 – 28, 2003.

[77] H. Wu, R. Xu, N. Harpaz, D. Burstein, and J. Gil. Segmentation of intestinal gland images with iterative region growing. *Journal of Microscopy*, 220(Part 3):190–204, 2005.

[78] H. Wu, R. Xu, N. Harpaz, D. Burstein, and J. Gil. Segmentation of microscopic images of small intestinal glands with directional 2-D filters. *Analytical and Quantitative Cytology and Histology*, 27(5):291–300, 2005.

[79] J. Yao, Z. M. Zhang, S. Antani, R. Long, and G. Thoma. Automatic medical image annotation and retrieval. *Neurocomputing*, 71(10-12):2012 – 2022, 2008.

[80] F. Yu and H. H. Ip. Semantic content analysis and annotation of histological images. *Computers in Biology and Medicine*, 38(6):635 – 649, 2008.

[81] K. R. Zalik. An efficient k'-means clustering algorithm. *Pattern Recognition Letters*, 29(9):1385 – 1391, 2008.

[82] D. Zhang, A. Wong, M. Indrawan, and G. Lu. Content-based image retrieval using gabor texture features. In *IEEE Transactions on PAMI*, pages 13–15, 2000.

[83] G. Zhang and Z.-M. Ma. Texture feature extraction and description using Gabor wavelet in content-based medical image retrieval. In *International Conference on Wavelet Analysis and Pattern Recognition*, volume 1-4, pages 169–173, 2007.

[84] Y.-F. Zhang, D.-L. Wu, M. Guan, W.-W. Liu, Z. Wu, Y.-M. Chen, W.-Z. Zhang, and Y. Lu. Tree analysis of mass spectral urine profiles discriminates transitional cell carcinoma of the bladder from noncancer patient. *Clinical Biochemistry*, 37(9):772 – 779, 2004.

[85] L. Zheng, A. Wetzel, J. Gilbertson, and M. Becich. Design and analysis of a content-based pathology image retrieval system. *IEEE Transactions on Information Technology in Biomedicine*, 7(4):249 –255, 2003.