Color Graphs for Automated Cancer Diagnosis and Grading

Dogan Altunbay, Celal Cigir, Cenk Sokmensuer, and Cigdem Gunduz-Demir*, *Member, IEEE*

Abstract—This paper reports a new structural method to mathematically represent and quantify a tissue for the purpose of automated and objective cancer diagnosis and grading. Unlike the previous structural methods, which quantify a tissue considering the spatial distributions of its cell nuclei, the proposed method relies on the use of distributions of multiple tissue components for the representation. To this end, it constructs a graph on multiple tissue components and colors its edges depending on the component types of their endpoints. Subsequently, it extracts a new set of structural features from these *color graphs* and uses these features in the classification of tissues. Working with the images of colon tissues, our experiments demonstrate that the color-graph approach leads to 82.65% test accuracy and that it significantly improves the performance of its counterparts.

Index Terms—Biomedical image processing, cancer, graph theory, histopathological image analysis, image representations, medical diagnosis.

I. INTRODUCTION

IN the human body, tissues are characterized with the organization of their components. Neoplastic diseases including cancer cause changes in these organizations. In order to diagnose and grade a neoplastic disease, pathologists visually examine the existence of such changes in the tissue. However, this examination mainly relies on the visual interpretation, and thus, it may lead to a considerable amount of intra- and interobserver variability, especially in cancer grading [1]. To reduce this observer variability, many studies have proposed to use computational methods that quantify the organizational properties of tissues.

In literature, there are mainly four different approaches for tissue quantification. These are *morphological*, *intensity-based*, *textural*, and *structural* approaches. In the morphological approach, a tissue is quantified with the size and shape properties of its cellular components. These properties are measured defining morphological features such as area, perimeter, roundness, and symmetry [2]. This approach requires identifying the exact boundaries of cells before extracting the features. However, due to the the complex nature of histopathological images, this gives rise to a difficult segmentation problem. In the intensity-based approach, a tissue is quantified with the gray level or color intensities of its pixels. This approach calculates a histogram of the tissue image by putting its pixels into bins and defines features such as average, standard deviation, and entropy on the histogram [3]. The intensity distributions are, however, similar for the images of tissues that are stained with the routinely used hematoxylin-and-eosin technique. This results in obtaining similar color distributions for different types of tissues. In the textural approach, the texture of the entire tissue is characterized with a set of features such as those calculated from co-occurrence matrices [4]–[6], run-length matrices [7], and multiwavelet coefficients [7], [8]. As this approach defines textures on pixels, it is sensitive to noise in the pixel values. In a typical histopathological image, there could be a large amount of noise arising from the staining and sectioning related problems [9].

In the structural approach, a tissue is characterized with the spatial distribution of its cellular components. To this end, a tissue is represented as a graph and a set of structural features is extracted from this graph representation. Previous studies have proposed to use different graph generation methods. Among these studies, a large set of them employs Delaunay triangulations (and their corresponding Voronoi diagrams) for tissue representation [4], [7], [10]–[14]. These studies define Delaunay triangulations on the centroids of the cell nuclei. In addition, they use minimum spanning trees that are constructed by connecting the centroids of cell nuclei based on their Euclidean distance [4], [7], [10]–[15]. Another set of the previous studies employs probabilistic graphs to represent a tissue [16]–[18]. In these studies, graphs are constructed on nuclear nodes by probabilistically assigning an edge between every pair of nodes; the probability of assigning an edge increases with the decreasing Euclidean distance between the end nodes of the edge. In [19], a weighted complete graph is generated to represent a tissue in which the edge weight is defined based on the Euclidean distance between the nodes.

All these structural studies consider the spatial locations of cell nuclei to represent a tissue, ignoring the existence of different components (such as luminal and stromal regions) in the tissue. Nevertheless, these components help better represent a tissue for some tissue types, and hence, they help better quantify the tissue changes. For example, in colon tissues, epithelial cells are lined up around a lumen to form glandular structures. In Fig. 1(a) and (b), the gland architecture for a normal tissue is shown (the cellular, stromal, and luminal components are indicated on an image in Fig. 2). Colon adenocarcinoma,[1] which

D. Altunbay and C. Cigir are with the Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey (e-mail: altunbay@cs.bilkent.edu.tr; cigir@cs.bilkent.edu.tr).

C. Sokmensuer is with the Department of Pathology, Medical School, Hacettepe University, Ankara TR-06100, Turkey (e-mail: csokmens@hacettepe.edu.tr).

*C. Gunduz-Demir is with the Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey (e-mail: gunduz@cs.bilkent.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2009.2033804

[1]Colon adenocarcinoma accounts for about 90–95 common cancer type in both men and women in North America and Western Europe [20].
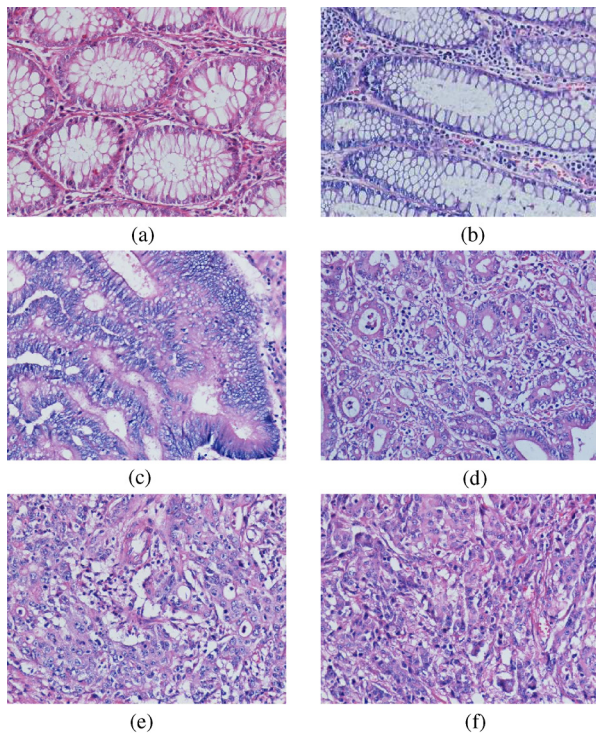
Fig. 1.   Histopathological images of colon tissues, which are stained with the routinely used hematoxylin-and-eosin technique. (a) and (b) Normal. (c) and (d) Low-grade cancerous. (e) and (f) High-grade cancerous.
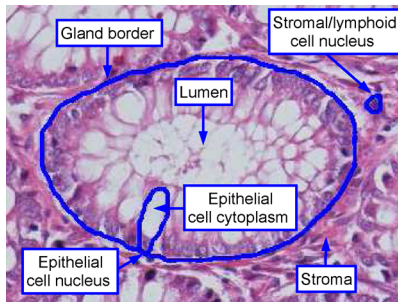

Fig. 2.   Cellular, stromal, and luminal components of a colon tissue.

originates from the epithelial cells, causes changes in the organizations of cells, leading to distortions in glandular structures. At the beginning, the degree of distortions is lower such that the gland formations are well to moderately differentiated; examples of such low-grade cancerous tissues are shown in Fig. 1(c) and (d). Then the distortion level becomes higher such that the gland formations are only poorly differentiated; examples of such high-grade cancerous tissues are shown in Fig. 1(e) and (f). The quantification of the distortions in glandular structures is very important in the correct identification of the cancer grade, which is one of the most important factors for the development of a cancer treatment plan. In addition to its cellular/nuclear components, a gland is characterized also with its luminal and stromal regions, as observed in Fig. 1. Thus, the additional information about the locations of luminal and stromal regions facilitates better quantification of the distortions.

This paper introduces a new structural approach that considers the locations of nuclear, stromal, and luminal components for tissue representation. For this purpose, it constructs a Delaunay triangulation on these tissue components and then colors each triangle edge based on the component types of its endpoints. From the constructed "color graph," the proposed approach extracts a set of structural features and uses these features for the automated cancer diagnosis and grading. Working with 213 photomicrographs of colon tissues that are taken from 58 different patients, our experiments demonstrate that the proposed color-graph approach leads to 82.65% test accuracy and that it significantly improves the classification performance of its counterparts. To the best of our knowledge, this is the first proposal for the use of color graphs to extract organizational principles of tissues for the purpose of cancer diagnosis and grading.

II. COLOR-GRAPH CONSTRUCTION

The proposed approach constructs a color graph for a tissue image by considering the centroids of tissue components as nodes and assigning color edges between these nodes making use of Delaunay triangulation. The following sections provide the details of node identification and edge assignment algorithms.

A. Node Identification

In the color-graph approach, nodes correspond to nuclear, stromal, and luminal tissue components. In the ideal way, one should segment each component, identifying its exact boundary. However, because of the complex nature of a histopathological image scene, this leads to a difficult segmentation problem even for the human eye. In a typical histopathological image, there could be staining and sectioning related problems, including the existence of touching and overlapping components, lack of dark separation lines between a component and its surroundings, inhomogeneity of the interior of a component, and presence of stain artifacts in a tissue [9]. Therefore, instead of determining their exact locations, the proposed approach approximately represents the tissue components with a set of circular primitives; the centroids of these tissue components are considered as the node locations in constructing a color graph. This representation has three types of circular primitives: one for nuclear components, one for stromal components, and one for luminal components.

In the definition of the circular primitives, the pixels of an image are first quantized into three clusters with the k-means algorithm. The number of clusters is particularly selected as three since there are mainly three color groups in the histopathological image of a tissue that is stained with the hematoxylin-and-eosin technique, which is routinely used to stain tissues in hospitals. These colors are purple, pink, and white, which typically correspond to nuclear, stromal, and luminal components, respectively. At the end of the k-means algorithm, the pixels are automatically labeled according to the intensity of their clustering vectors; the cluster vector with the highest intensity value and its corresponding pixels are labeled as white (which are
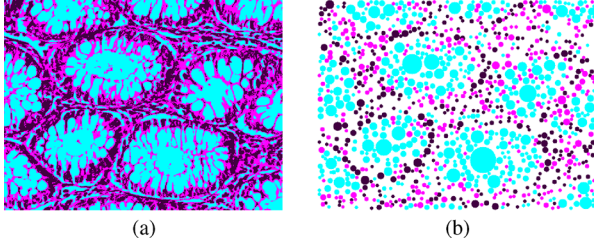
Fig. 3. (a) Pixels after color quantization and (b) circular primitives located on these quantized pixels. Here, purple, pink, and cyan correspond to nuclear, stromal, and luminal components, respectively.

TABLE I
LIST OF COLORS THAT ARE USED TO COLOR TRIANGLE EDGES

Red	: Luminal-luminal edge
Yellow	: Stromal-stromal edge
Black	: Nuclear-nuclear edge
Blue	: Luminal-stromal edge
Green	: Luminal-nuclear edge
Magenta	: Stromal-nuclear edge

shown with cyan in Fig. 3), the cluster vector with the lowest intensity value and its pixels are labeled as purple, and the remaining one and its pixels are labeled as pink.

On the pixels of each cluster, morphological operators (closing followed by opening with a square structuring element of the size of 3) are first applied to reduce the noise that arises from the incorrect clustering of pixels. Subsequently, a set of circles is located on these pixels by making use of the heuristic algorithm proposed in [21]. In this algorithm, circles are iteratively located on a given set of pixels, provided that the area of these circles is greater than an area threshold. In this study, the area threshold is selected as 25, considering the resolution of images and the magnification of the microscope objective lens that is used to take the images; this area threshold corresponds to eliminating circles with a radius smaller than 3. Note that as the resolution and/or the magnification increases, larger values of this threshold should be selected. In this study, a circular shape is selected for the representation since tissue components typically have curvy boundaries and since circles are efficiently located on a set of pixels. Fig. 3(a) illustrates the clustered pixels of the image that is given in Fig. 1(a); in this figure, nuclear, stromal, and luminal components are shown with purple, pink, and cyan, respectively. Fig. 3(b) shows the circles located on the clustered pixels; similarly, circles corresponding to nuclear, stromal, and luminal components are shown with purple, pink, and cyan, respectively.

B. Edge Assignment

After identifying the centroids of the circular primitives as nodes, the proposed algorithm assigns edges between these nodes. For this purpose, it first constructs a Delaunay triangulation on the graph nodes. It then colors each triangle edge depending on the component types of its end nodes. As there are three different types of components, edges are colored with one of the six colors that are given in Table I. For the image given in Fig. 1(a), the edges assigned using this algorithm are illustrated
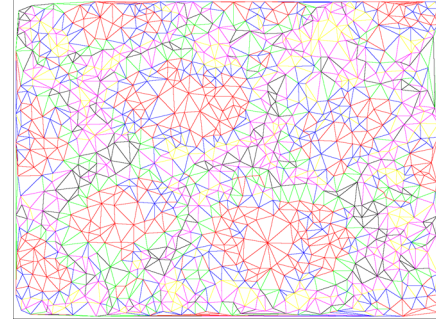

Fig. 4. Edges that are obtained by using Delaunay triangulation and that are colored with one of the six colors depending on the component types of their end nodes.

in Fig. 4. Note that our implementation employs a MATLAB function that uses the Qhull algorithm [22] for computing Delaunay triangulation.

III. COLOR-GRAPH FEATURES

In order to quantify a color graph, three global graph properties are used. These properties are the average degree, the average clustering coefficient, and the diameter. The definition of these properties is given in the following sections.

A. Average Degree

The degree of a node is defined as the number of its edges. The average degree of a graph is the mean of the degrees computed for every node in the graph. In this study, for a single node, seven different degrees are defined. The first one is computed considering all of its edges, regardless of their colors. The next six of them are computed considering the edge colors; for a particular color, a node degree is defined as the number of edges that belong to the corresponding node and that are of this particular color. Taking their average, seven average degrees are obtained for an individual color graph.

B. Average Clustering Coefficient

The clustering coefficient of a node quantifies the connectivity information in its neighborhood [23]. The clustering coefficient CC_i for node i is defined as

$$CC_i = \frac{2E_i}{d_i(d_i - 1)} \qquad (1)$$

where d_i is the number of the neighbors of node i and E_i is the number of the existing edges between these neighbors. For a color graph, four average clustering coefficients are defined. The first one is the average of the clustering coefficients of nodes that are computed using (1), without considering the edge colors. Here, the values of nodes with a degree of 0 or 1 are excluded from the average computation as these degrees lead to indeterminate quotients.

The additional three average clustering coefficients are defined to quantify the clustering information of nodes of the same component type. For this purpose, for each node, a colored clustering coefficient is defined considering only the corresponding

edges. For instance, for a luminal component node, only the luminal-luminal (red) edges are considered and a red clustering coefficient is defined as

$$CC_i^{(\text{red})} = \frac{2E_i^{(\text{red})}}{d_i^{(\text{red})}(d_i^{(\text{red})} - 1)} \qquad (2)$$

where $d_i^{(\text{red})}$ is the number of the luminal neighbors of node i and $E_i^{(\text{red})}$ is the number of the existing red edges between these neighbors. For nuclear and stromal components, clustering coefficients are computed in a similar way. Taking the average of these colored clustering coefficients, three more features are obtained for an individual color graph.[2]

C. Diameter

The diameter of a graph is the longest of the shortest paths between any pair of graph nodes. In this study, seven diameters for a color graph are computed. Likewise, the first diameter is computed without considering the edge colors and the other six of them are computed considering only the edges with a color of interest. For instance, the red diameter is the longest of the shortest paths between any pair of nodes that are reachable to each other using only the red edges.

IV. EXPERIMENTS

We conduct our experiments on 213 microscopic images of colon tissues taken from 58 different patients; these patients are randomly chosen from the Pathology Department archives in Hacettepe University School of Medicine. These tissue samples are stained with hematoxylin-and-eosin and consist of 5-μm-thick tissue sections.

In this study, the images are assumed to be homogeneous, similar to most of the previous studies. Therefore, in a fully automated system, unsupervised segmentation is necessary beforehand. In literature, there are different unsupervised segmentation algorithms that are proposed to divide a heterogeneous image into its homogenous regions [26]–[28]. In our previous work [21], we propose such an algorithm that is specifically designed for the unsupervised segmentation of tissue images. That algorithm defines a homogeneity criterion, which is used in segmentation, based on the spatial distribution and size variety of circular primitives.

The dataset includes the images of normal, low-grade adenocarcinomatous, and high-grade adenocarcinomatous colon tissues. Each tissue image is labeled by our medical expert.[3]

---

[2]In the definition of the clustering coefficient, we only consider the colors that correspond to edges with the components of the same type but not to edges with the cross components. If the clustering coefficient was calculated for the colors corresponding to cross components, its value would always be zero. For example, suppose that the blue clustering coefficient is computed for a luminal node. This color is defined for edges between luminal and stromal components. Hence, the neighbors that are to be considered in the clustering coefficient calculation will be stromal components. However, between stromal components, it is not possible to define blue edges.

[3]By Prof. C. Sokmensuer, MD, who is specialized in colorectal carcinomas and has been practicing oncology for the past 15 years. Each image is shown to him for three different times and at the end, he decides on the image label considering his three decisions.

TABLE II
CONFUSION MATRICES OBTAINED WITH COLOR-GRAPH APPROACH FOR THE
(a) TRAINING SET AND (b) TEST SET (THESE RESULTS ARE OBTAINED WHEN
ALL COLOR-GRAPH FEATURES ARE USED)

			Computed class		
			Normal	Low-grade	High-grade
Actual	class	Normal	35	2	1
		Low-grade	0	31	6
		High-grade	0	4	36

(a)

			Computed class		
			Normal	Low-grade	High-grade
Actual	class	Normal	29	2	3
		Low-grade	3	27	5
		High-grade	2	2	25

(b)

The dataset is randomly divided into training and test sets. The training set consists of 115 images of tissues (38 normal, 37 low-grade cancerous, and 40 high-grade cancerous) taken from 29 patients. The test set consists of 98 images of tissues (34 normal, 35 low-grade cancerous, and 29 high-grade cancerous) taken from the remaining 29 patients.

The tissue images that are used in the experiments are taken using a Nikon Coolscope Digital Microscope with $20\times$ microscope objective lens. This magnification is high enough to obtain homogeneous images and at the same time low enough to obtain images containing multiple glands. The images are taken in the RGB color space and then converted to the Lab color space for further processing.

The image resolution affects the computational time required for processing a single image. Thus, the image resolution is selected as 480×640, which gives both accurate classification results and relatively lower computational times. The computational time required for processing a single image is $6.47(\pm 0.62)$ seconds on the average. These results are obtained on a computer with a 2.83-GHz processor and 4 GB of RAM.

For classifying the tissue images, we use support vector machines (SVMs) with linear kernels. Here, we use the implementation that is available at http://svmlight.joachims.org [24]; this is an implementation of the multiclass SVMs described in [25]. We set all of its learning parameters to their default values except the value of parameter C, which affects the tradeoff between the training error and the margin. In the experiments, its value is selected using ten-fold cross-validation on the training set; note that none of the samples in the test set is used in this selection. In particular, a candidate set of $\{1, 2, \ldots, 9, 10, 20, \ldots, 90, 100, 150, \ldots, 950, 1000\}$ is considered for C and the one that leads the best cross-validation accuracy is selected.

V. RESULTS AND DISCUSSION

In Table II, we provide the confusion matrices obtained on the training and test sets. These results are obtained selecting the SVM parameter C as 500 using ten-fold cross validation. Table II shows that the color-graph approach yields high accuracies for both the training and test sets. It also shows that although the proposed approach leads to higher accuracies for
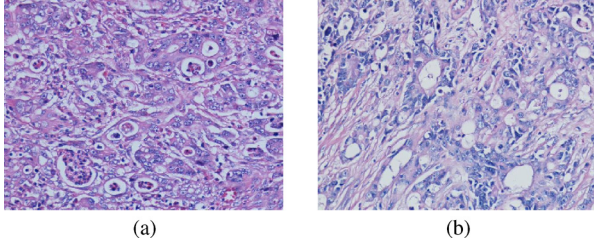
Fig. 5. Examples of the low-grade cancerous tissues that are incorrectly classified as high-grade cancerous by the color-graph approach. Note that these tissues lie at the boundary between low-grade and high-grade cancer.

TABLE III
AVERAGE ACCURACY RESULTS AND THEIR STANDARD DEVIATIONS OBTAINED
BY APPLYING LEAVE-ONE-PATIENT-OUT CROSS-VALIDATION ON THE TEST SET

	All features	Selected features
Normal	85.29±1.39	85.29±1.39
Low-grade	77.16±1.81	80.01±1.54
High-grade	86.19±0.81	82.73±1.42
Overall	82.65±0.85	82.65±0.86

The results are obtained when all color-graph features are used and when the seven features selected by backward elimination are used.

TABLE IV
LIST OF FEATURES SELECTED BY BACKWARD ELIMINATION

$Degree(yellow)$	The average degree for stromal-stromal edges
$Degree(blue)$	The average degree for luminal-stromal edges
$Degree(green)$	The average degree for luminal-nuclear edges
$CC(red)$	The average clustering coefficient for luminal components
$CC(black)$	The average clustering coefficient for nuclear components
$Diameter(yellow)$	The diameter for stromal-stromal edges
$Diameter(blue)$	The diameter for luminal-stromal edges

TABLE V
CONFUSION MATRICES OBTAINED WITH THE COLOR-GRAPH APPROACH FOR
THE (a) TRAINING SET AND (b) TEST SET

			Computed class		
			Normal	Low-grade	High-grade
Actual class		Normal	32	4	2
		Low-grade	0	32	5
		High-grade	0	3	37

(a)

			Computed class		
			Normal	Low-grade	High-grade
Actual class		Normal	29	2	3
		Low-grade	3	28	4
		High-grade	0	5	24

(b)

These results are obtained when seven color-graph features that are selected using backward elimination are used.

the classification of the normal and high-grade cancer, it gives relatively lower accuracies for the low-grade cancer. For this reason, we check the misclassified low-grade cancerous tissues. Here, we observe that almost half of these tissues are the cases lying at the boundary between low-grade and high-grade cancer and that these cases are incorrectly classified as high-grade cancerous by the proposed approach. In Fig. 5, two examples of such boundary cases are illustrated.

In order to examine how much stable the results are across multiple images (i.e., how much the classification results are deviated with different sets of data), we apply "leave-one-patient-out" cross-validation. To this end, from the test data, we obtain multiple sets in each of which we exclude the samples of a particular patient. The average results obtained by this cross validation and their standard deviations are given in the first column of Table III. This table shows that the results are stable and do not show too much differences with the inclusion/exclusion of a particular patient.

A. Feature Selection

In the proposed approach, 18 color-graph features are defined and used for tissue classification. Nevertheless, this number of features is relatively large given the data size. As this might cause an over-fitting problem, we select seven features by using backward elimination. At each step of this procedure, the feature exclusion of which gives the highest training accuracy is removed. For each feature, a SVM classifier is used and its parameter C is selected using ten-fold cross-validation on the training set. Note that, in this procedure, test samples are not used at all.

The selected features are listed in Table IV. Using the set of these features, the confusion matrices obtained for the training and test sets are given in Table V. Similarly, we apply leave-one-patient-out cross validation to examine how much the results are deviated with the use of different datasets. The results of this

cross validation are reported in the second column of Table III. All these tables show that there exist some redundant features and that similar accuracy results are obtained with a reduced set of features.

Next, the selected features are further examined to understand their underlying biological phenomena. The intuitions based on medical knowledge and the observations made during experiments are summarized as follows. In *a normal colon tissue*, epithelial cells are arranged in a mono-layer structure around a lumen to form a glandular region. When hematoxylin-and-eosin staining is used, the nuclei of these cells appear in purple-like color and lumens appear in white. The epithelial cells are rich in mucin, and hence, their cytoplasms appear in very light color. In our approach, these cytoplasms are also represented with white circles (for simplicity, all white circles are referred as luminal components). Since the epithelial cell cytoplasms and luminal regions, and thus their corresponding white circles, are in very close proximity in a normal tissue, other types of circles (components) rarely exist in between the corresponding white circles. As a result, the average clustering coefficient of luminal components (CC(red)) is expected to be higher. At the same time, since the epithelial cells are arranged in a mono-layer structure, it is expected to have ring-like connections between their corresponding purple circles. This results in having lower clustering coefficients of nuclear components (CC(black)).

In *a low-grade adenocarcinomatous tissue*, the epithelial cells become to be poor in mucin, and hence, their cytoplasms appear in pink-like color. These cytoplasms are represented with a group of pink circles (for simplicity, all pink circles are referred as stromal components). Since the epithelial cells (also their cytoplasms) are close to each other, it could be expected to have more edges between the corresponding pink circles (stromal

components). This is attributed to the increase in the degree and the diameter for stromal-stromal edges (degree(yellow) and diameter(yellow)). Additionally, in some of these tissues, the epithelial cells begin to aggregate, which causes for them to be arranged in a multilayer structure around a lumen. This multilayer structure leads to more purple circles (nuclear components) being connected to each other, which increases the average clustering coefficient of nuclear components (CC(black)).

In *a high-grade adenocarcinomatous tissue*, the distortion becomes higher, causing much more chaotic distribution of tissue components. The tissue components, and their corresponding circles, seem to randomly distribute all over the tissue without forming differentiated glands. The randomness in the tissue (the absence of glands) causes more variety in connections between different types of components whereas the existence of glands may favor some connection types that are previously mentioned. In our experiments, we observe that the sum of luminal-stromal edges and luminal-nuclear edges per each corresponding node increases (the sum of degree(blue) and degree(green)). This is attributed to the randomness in high-grade cancerous tissues.

B. Analysis of Node Identification

The first step of node identification is to cluster the pixels of tissue images into three. As they are prepared in different times and by different people, there could exist staining differences between the tissue samples. However, the pixels of each image are separately clustered and then labeled according to the intensity of their clustering vectors.

In its next step, tissue components (nodes) are segmented on the clustered pixels. In the color-graph approach, segmentation relies on the approximate representation of tissue components. As this segmentation does not exactly locate the components in the tissue, there could be some information lost. Moreover, as there could be a considerable amount of noise arising from the tissue preparation procedures, some tissue components could be located incorrectly. To understand the effectiveness of the proposed segmentation method, we quantitatively assess the effects of incorrect segmentation to the classification accuracy. To this end, for a given tissue image, we make modifications on the m percent of its segmented nodes. These modifications include a randomly selected combination of the following operations.

1) *Insertion*: This operation adds a new node into the tissue representation. The coordinates of the node and its tissue type are randomly determined. The aim of the insertion operation is to simulate the cases where additional circles are incorrectly identified because of the noise and artifacts in the tissue as well as the segmentation algorithm (e.g., a purple circle located on a stain artifact).

2) *Deletion*: This operation removes an existing node from the tissue representation. The node, which is to be removed, is randomly selected. The aim of the deletion operation is to simulate the cases where some nodes cannot be identified because of the similar reasons (e.g., a nucleus that cannot be represented with a purple circle due to a fading in stain).
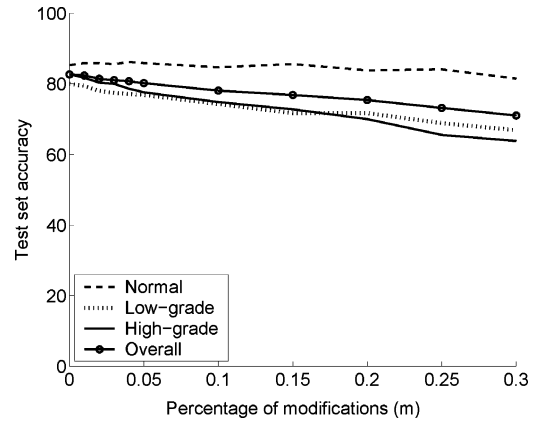

Fig. 6.   Effects of the changes in node locations to the test set accuracy. These results are obtained when seven color graph features that are selected using backward elimination are used.

3) *Type change*: This operation changes the tissue type of an existing node. The node whose type is to be changed and the new type are randomly selected. The aim of this operation is to simulate the cases where stain artifacts exist in the tissue and/or the k-means algorithm results in incorrect clustering of some pixels (e.g., a cell nucleus appearing in pink because of stain artifacts).

Fig. 6 shows the accuracy as a function of m, which is the percentage of modifications. These accuracy results are obtained on the test set when the selected features are used in classification. Fig. 6 shows that the proposed method still leads to accurate results ($>80\%$) when the value of m is smaller than or equal to 0.05; note that the number of modifications is $58.60(\pm 4.11)$ on the average when $m = 0.05$. Here, the error arising from the modified nodes could most probably be compensated by the other nodes. With the increasing number of modifications, the accuracy of classifying low-grade and high-grade cancerous tissues decreases and the overall accuracy reaches down to approximately 70% when $m = 0.3$. These results show that the proposed approach provides a tolerance to slight differences in segmentation.

C. Parameter Analysis

Besides SVM parameter C, there are two external parameters, which are used in node identification. These are: 1) *the area threshold* that is used to locate circular primitives and 2) *the size of the structuring element* that is used by morphological operators. To investigate the effects of these parameters to the classification accuracy, we fix one of the parameters and observe the accuracy as a function of the other parameter. Fig. 7(a) gives the accuracies obtained on the test set when the area threshold is selected as {5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, 150}. This figure shows that there is a range for this parameter to give accurate results (values ≥ 15 and ≤ 45). When the threshold becomes higher (values >50), there is a huge decrease in the classification accuracy of high-grade cancerous tissues.

Similarly, Fig. 7(b) shows the accuracies obtained on the test set when the structuring element size is selected as {1, 3, 5, 7, 9, 11, 13}; here the value of 1 corresponds to the case where no
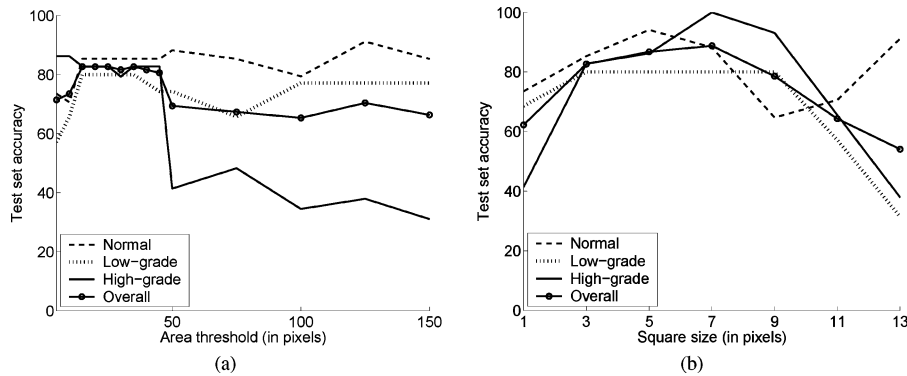
Fig. 7. Effects of the selection of the parameters in the node identification step to the classification accuracy: (a) area threshold and (b) size of the structuring element. These results are obtained when seven color-graph features that are selected using backward elimination are used.

TABLE VI
CLASSIFICATION ACCURACIES OBTAINED BY COLOR-GRAPH APPROACH AS WELL AS BY INTENSITY-BASED, TEXTURAL,
AND OTHER STRUCTURAL APPROACHES

| | Training accuracies | | | | Test accuracies | | | | Number of |
	Normal	Low-grade	High-grade	Overall	Normal	Low-grade	High-grade	Overall	dimensions
Color graphs (all features)	92.11	83.78	90.00	88.70	85.29	77.14	86.21	82.65	18
Color graphs (selected features)	84.21	86.49	92.50	87.83	85.29	80.00	82.76	82.65	7
Gray-level histograms	78.95	78.38	30.00	61.74	85.29	77.14	0.00	57.14	4
Co-occurrence matrices	73.68	51.35	82.50	69.57	61.76	68.57	37.93	57.14	8
Run-length matrices	89.47	35.14	90.00	72.17	79.41	14.29	72.41	54.08	7
Colorless graphs	50.00	32.43	65.00	49.57	50.00	37.14	58.62	47.96	3
Delaunay triangulations	84.21	51.35	77.50	71.30	76.47	57.14	72.41	68.37	11
Probabilistic graphs	45.26	43.24	53.25	47.39	40.00	57.43	34.48	44.59	7

morphological operator is applied. These results show that it is necessary to apply morphological operators for noise reduction and that the structuring element sizes in between 3 and 7 give the most accurate results. In the analysis of both of these parameters, SVM parameter C is selected using ten-fold cross-validation on the training set.

VI. COMPARISONS WITH OTHER APPROACHES

To investigate the effectiveness of the color-graph representation, we compare its results against those of the others that use intensity-based, textural, and structural features. Here, we exclude the use of morphological features from the comparisons since this requires the correct segmentation of cells, which is beyond the scope of this paper. In all comparisons, we also use SVMs (with linear kernels) as classifiers. Similarly, SVM parameter C is selected from the candidate set of $\{1, 2, \ldots, 9, 10, 20, \ldots, 90, 100, 150, \ldots, 950, 1000\}$ using ten-fold cross-validation on the training set.

A. Comparison With Intensity-Based Features

In the intensity-based approach, features are extracted from the gray-level intensity histograms. For this purpose, gray-level pixels are put into N bins and the average, standard deviation, skewness, and kurtosis are computed on these bin values [3]. In our experiments, we consider the values of 8, 16, and 32 for N and the same candidate set for SVM parameter C. Evaluating ten-fold cross-validation performance for all possible combinations of these two parameters, we select N and C as 8 and 900, respectively. Table VI reports the training and test set accura-

cies obtained by the intensity-based representation. This table shows that this representation yields lower accuracies, for especially high-grade cancerous tissues. In order to understand the statistical significance of these results, we apply the McNemar's test to the 2×2 contingency table. With a significance level of 0.05, it shows that there is a statistically significant difference between the accuracies of the color-graph and intensity-based approaches.

B. Comparison With Textural Features

In our experiments, we use two sets of textural features that are extracted from co-occurrence and run length matrices.

1) Co-occurrence Matrices: The co-occurrence matrix $M_{d,\theta}$ is the second order histogram of the pairwise pixel values. Each of its entry, $M_{d,\theta}(i, j)$, keeps the number of times that gray value i co-occurs with gray value j in a particular spatial relationship defined by distance d and angle θ. For a given distance d, we use eight co-occurrence matrices each of which is computed with one of the following angles $\theta = \{0, \pi/4, 2\pi/4, 3\pi/4, 4\pi/4, 5\pi/4, 6\pi/4, 7\pi/4\}$. Then, we compute matrix M_d as the average of these eight matrices since the characteristics of a histopathological image is rotation invariant. After normalizing the average matrix, we extract eight textural features [29]; these features are listed in Table VII. In this table, μ_x and μ_y are the means of row and column sums, respectively, and σ_x and σ_y are their standard deviations.

In the computation of a co-occurrence matrix, gray values are usually quantized into bins to alleviate the effects of the noise in pixel values. Thus, this approach has three model parameters: the

TABLE VII
TEXTURAL FEATURES DERIVED FROM CO-OCCURRENCE MATRIX

$$\text{energy} = \sum_i \sum_j M_d(i,j)^2$$

$$\text{entropy} = -\sum_i \sum_j M_d(i,j) \log M_d(i,j)$$

$$\text{contrast} = \sum_i \sum_j (i-j)^2 M_d(i,j)$$

$$\text{homogeneity} = \sum_i \sum_j \frac{M_d(i,j)}{1+|i-j|}$$

$$\text{correlation} = \sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)M_d(i,j)}{\sigma_x \sigma_y}$$

$$\text{dissimilarity} = \sum_i \sum_j |i-j| M_d(i,j)$$

$$\text{inverse difference moment} = \sum_i \sum_j \frac{M_d(i,j)}{1+(i-j)^2}$$

$$\text{maximum probability} = \max_{i,j} M_d(i,j)$$

TABLE VIII
TEXTURAL FEATURES DERIVED FROM RUN-LENGTH MATRIX

$$\text{short run emphasis} = \frac{\sum_i \sum_j \frac{R(i,j)}{j^2}}{n}$$

$$\text{long run emphasis} = \frac{\sum_i \sum_j R(i,j) \cdot j^2}{n}$$

$$\text{gray level nonuniformity} = \frac{\sum_i (\sum_j R(i,j))^2}{n}$$

$$\text{run length nonuniformity} = \frac{\sum_j (\sum_i R(i,j))^2}{n}$$

$$\text{run percentage} = \frac{\sum_i \sum_j R(i,j)}{p}$$

$$\text{low gray level runs emphasis} = \frac{\sum_i \sum_j \frac{R(i,j)}{i^2}}{n}$$

$$\text{high gray level runs emphasis} = \frac{\sum_i \sum_j R(i,j) \cdot i^2}{n}$$
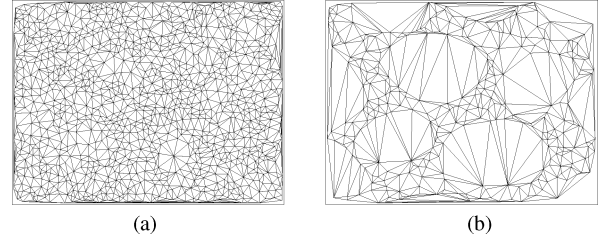

Fig. 8. (a) Colorless edges that are obtained using Delaunay triangulation constructed on nuclear, stromal, and luminal tissue components and (b) colorless edges that are obtained by constructing a Delaunay triangulation on nuclear components.

number of bins N, the value of distance d, and SVM parameter C. In our experiments, we consider the set of $\{8, 16, 32\}$ for N, the set of $\{5, 10, 20, 40\}$ for d, and the same candidate set for C. Considering the results of ten-fold cross-validation for all possible combinations of these three parameter values, we set N, d, and C to 8, 20, and 1, respectively. The training and test set accuracies obtained using this parameter set are also given in Table VI. These accuracies show that the features extracted from the co-occurrence matrix are not sufficient for successful classification of tissues. The McNemar's test with a significance level of 0.05 shows that the difference between these accuracies and those of the color-graph approach is statistically significant.

2) Run-Length Matrices: The run-length matrix R_θ defines higher order statistical features. Each of its entry, $R_\theta(i,j)$, keeps the number of runs (a set of consecutive, collinear pixels) of gray values i with a length of j in the direction of θ. In this representation, we compute four different run-length matrices at four different angles $\theta = \{0, \pi/4, 2\pi/4, 3\pi/4\}$. Then, we calculate matrix R as the average of these four matrices and extract the seven features [30], [31]; these features are listed in Table VIII. In this table, n is defined as $\sum_i \sum_j R(i,j)$ and p is the number of pixels in the image.

Similarly, pixels are quantized into N bins before computing the run-length matrices. Thus, this approach has two model parameters: the number of bins N and SVM parameter C. In our experiments, we consider the values of $\{8, 16, 32\}$ for N and the same set of values for C. From all possible pairs, we select N and C as 8 and 900 using ten-fold cross-validation. The accuracy results obtained on the training and test sets are also given in Table VI. These accuracies are also lower than those of the color-graph approach; the McNemar's test shows that this is statistically significant.

C. Comparison With Structural Features

In the experiments, three colorless structural representations are used for comparison. Each of these representations uses a different method for graph generation.

1) Colorless Graphs: The colorless-graph representation also constructs a Delaunay triangulation considering the locations of nuclear, stromal, and luminal tissue components. However, it does not color the triangle edges, and thus, it does not use the colored versions of graph features. For the tissue image given in Fig. 1(a), the colorless edges obtained by this representation are shown in Fig. 8(a).

This representation does not have any parameter except SVM parameter C. We select its value as 2 making use of ten-fold cross-validation. Table VI reports the training and test set accuracies obtained using the colorless graphs. These results show that the exclusion of color information decreases the classification accuracy. The McNemar's test shows that the difference between accuracies of the color graphs and colorless graphs is statistically significant with a significance level of 0.05.

2) Delaunay Triangulation: In this representation, a Delaunay triangulation is constructed on nuclear tissue components; Fig. 8(b) illustrates the edges of such Delaunay triangulation for the image given in Fig. 1(a). As there is no color information associated with edges, the colorless versions of the features are used to quantify this representation, i.e., the average degree, the average clustering coefficient, and the diameter are computed without considering any edge colors. Furthermore, eight more features are extracted using the edge lengths and triangle areas [14]. For each of them, we compute the average, standard deviation, minimum-to-maximum ratio, and disorder. In the definition of disorder, we use the following formula that is given

in [12]

$$\text{disorder} = 1 - \frac{1}{1 + (\text{standard deviation/average})}. \quad (3)$$

As a result, a total of 11 features is used to quantify a Delaunay triangulation. Similarly, a SVM classifier is used; its parameter C is selected as 650 using ten-fold cross-validation. The accuracy results obtained by this Delaunay triangulation representation are given in Table VI. These results demonstrate that the use of the locations of additional tissue components increases the classification accuracy especially for the low-grade cancerous tissues. The McNemar's test with a significance level of 0.05 shows that this increase is statistically significant.

3) Probabilistic Graphs: In the probabilistic graph representation, an unweighted and undirected graph is generated considering nuclear tissue components as nodes and probabilistically assigning edges between each pair of these nodes. The probability of assigning an edge between nodes u and v is computed as $P(u, v) = d(u, v)^{-\alpha}$, where $d(u, v)$ is the Euclidean distance between nodes u and v, and α is the exponent that should be selected greater than 0. More specifically, an edge $e(u, v)$ is defined between nodes u and v if $P(u, v) > r$, where r is a real number between 0 and 1 that is generated by a random number generator [17].

From these probabilistic graphs, seven topological features are extracted. The first three of them are the average degree, the average clustering coefficient, and the diameter. The last four additional features are: 1) the number of isolated nodes, which gives the number of nodes with a degree of 0; 2) the number of end nodes, which gives the number of nodes with a degree of 1; 3) the number of the connected components in a graph; and 4) the ratio of the giant connected component[4] size to that of the entire graph.

In addition to SVM parameter C, this representation has another model parameter α, which controls the density of graph edges. In our experiments, we use the same candidate set for parameter C and use a candidate set of $\{0.1, 0.2, \ldots, 0.9, 1.0\}$ for parameter α. We evaluate the ten-fold cross-validation performance for all possible combinations of these parameter sets. Considering these performances, we select the value of α as 0.1 and the value of C as 2. In this representation, we may have different graphs for the same tissue image as these graphs are probabilistically generated. Thus, we generate ten different graphs for each image in the dataset and select the parameters according to the average cross-validation performance obtained on these ten graphs. In Table VI, we report the accuracy of the probabilistic graph representation obtained on training and test sets; these are the average accuracies obtained on ten different sets of graphs. This table shows that this representation also leads to lower accuracy results compared to the color-graph approach. The McNemar's test with a significance level of 0.05 shows that these results are also statistically significant.

VII. CONCLUSION

In this paper, we report a new tissue representation in which the locations of different tissue components are considered. In this representation, a color graph is generated constructing a Delaunay triangulation on nuclear, stromal, and luminal tissue components and coloring the triangle edges according to the component types of their endpoints. On the generated color graph, the colored versions of the average degree, the average clustering coefficient, and the diameter are defined considering the edge colors.

In this study, we conduct the experiments on 213 photomicrographs of colon tissues that are taken from 58 different patients. These experiments demonstrate that the colored features lead to 82.65% test accuracy for the automated colon cancer diagnosis and grading. In this study, we also compare the color-graph approach with the intensity-based, textural, and colorless structural approaches. Our experiments also show that the proposed color-graph approach gives statistically better results compared to its counterparts.

In this study, we propose to use Delaunay triangulation for generating a graph since it is known to be one of the effective representations in quantifying the spatial distribution and since it does not require selecting an external parameter, for example, as in the case of probabilistic graphs in which parameter α should be selected. However, one could consider using different graph types. For instance, a probabilistic graph could be generated on tissue components and its edges could be colored similarly. Alternatively, a probabilistic graph could be constructed such that the distance between the segmented components and/or the size of these components are used in the coloring process. After constructing such probabilistic graphs, similar features could be used. It is also possible to employ Voronoi diagrams to represent a tissue. For example, the diagram could be colored according to the component type of its polygons. In this case, a set of color-polygon features could be extracted (e.g., the average area of polygons could be defined separately for each color). Alternatively, the polygon edges could be colored according to the type of components sharing these edges. All these aspects could be considered as the future research perspectives of using the coloring idea.

As another future research work, the effectiveness of the color-graph approach could be investigated for other types of tissues to capture structural and architectural differences observed in these tissues. Furthermore, one could also think of the use of color graphs (and color-graph features) for unsupervised segmentation of heterogeneous tissue images. Additionally, one would work on defining the eigenvalue spectrum of these color graphs and use the spectral features for classification.

---

[4]The giant connected component of a graph is its largest connected component, i.e., it is the largest set of nodes in which each node in this largest set is reachable from the every other node in the same set.

REFERENCES

[1] G. D. Thomas, M. F. Dixon, N. C. Smeeton, and N. S. Williams, "Observer variation in the histological grading of rectal carcinoma," *J. Clin. Pathol.*, vol. 36, pp. 385–391, 1983.

[2] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proc. Int. Symp. Electron. Imag.: Sci. Technol.*, 1993, vol. 1905, pp. 861–870.

[3] M. Wiltgen, A. Gerger, and J. Smolle, "Tissue counter analysis of benign common nevi and malignant melanoma," *Int. J. Med. Inf.*, vol. 69, pp. 17–28, 2003.

[4] 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 *Proc. Biomed. Imag.: From Nano Macro*, 2007, pp. 1284–1287.

[5] 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 Trans. Inf. Technol. B*, vol. 2, no. 3, pp. 197–203, Sep. 1998.

[6] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern Recognit.*, vol. 42, no. 6, pp. 1093–1103, 2009.

[7] B. Weyn, G. van de Wouver, M. Koprowski, A. van Daele, P. Scheunders, W. Jacob, and E. van Marck, "Value of morphometry, texture analysis, densitometry, and histometry in the differential diagnosis and prognosis of malignant mesothelia," *J. Pathol.*, vol. 189, pp. 581–589, 1999.

[8] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 697–704, Jun. 2003.

[9] J. Gil, H. Wu, and B. Y. Wang, "Image analysis and morphometry in the diagnosis of breast cancer," *Microsc. Res. Tech.*, vol. 59, pp. 109–118, 2002.

[10] 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*, vol. 35, pp. 23–29, 1999.

[11] S. J. Keenan, J. Diamond, W. G. McCluggage, H. Bharucha, D. Thompson, B. H. Bartels, and P. W. Hamilton, "An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)," *J. Pathol.*, vol. 192, pp. 351–362, 2000.

[12] J. Sudbo, R. Marcelpoil, and A. Reith, "New algorithms based on the Voronoi diagram applied in a pilot study on normal mucosa and carcinomas," *Anal. Cell. Pathol.*, vol. 21, pp. 71–86, 2000.

[13] M. Guillaud, D. Cox, K. Adler-Storthz, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekerk, N. Poulin, M. Follen, and C. MacAulay, "Exploratory analysis of quantitative histopathology of cervical intraepithelial neoplasia: Objectivity, reproducibility, malignancy-associated changes, and human papillomavirus," *Cytom. A*, vol. 60, no. 1, pp. 81–89, 2004.

[14] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," *Proc. Biomed. Imag.: From Nano Macro*, 2008, pp. 496–499.

[15] H.-K. Choi, T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P.-U. Malmstrom, and C. Busch, "Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility," *Anal. Cell. Pathol.*, vol. 15, pp. 1–18, 1997.

[16] C. Gunduz, B. Yener, and S. H. Gultekin, "The cell-graphs of cancer," *Bioinformatics*, vol. 20, no. 1, pp. i145–i151, 2004.

[17] C. Demir, S. H. Gultekin, and B. Yener, "Learning the topological properties of brain tumors," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 2, no. 3, pp. 262–270, Jul. 2005.

[18] C. Gunduz-Demir, "Mathematical modeling of the malignancy of cancer using graph evolution," *Math. Biosci.*, vol. 209, no. 2, pp. 514–527, 2007.

[19] C. Demir, S. H. Gultekin, and B. Yener, "Augmented cell-graphs for automated cancer diagnosis," *Bioinformatics*, vol. 21, no. 2, pp. ii7–ii12, 2005.

[20] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer Statistics 2008," *CA Cancer J. Clin.*, vol. 58, pp. 71–96, 2008.

[21] 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 Recognit.*, vol. 42, no. 6, pp. 1104–1112, 2009.

[22] C. B. Barber, D. P. Dobkin, and H. T. Huhdanpaa, "The Quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, 1996.

[23] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Adv. Phys.*, vol. 51, pp. 1079–1187, 2002.

[24] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[25] K. Cramme and Y. Singer, "On the algorithmic implementation of multi-class kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2002.

[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[27] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[28] J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz, "Adaptive perceptual color-texture image segmentation," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1524–1536, Oct. 2005.

[29] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.

[30] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Vis. Graph.*, vol. 4, pp. 172–179, 1975.

[31] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognit. Lett.*, vol. 11, pp. 415–420, 1990.

Dogan Altunbay received the B.S. degree in computer sciences and engineering from Hacettepe University, Ankara, Turkey, in 2007. He is currently working toward the M.S. degree in computer engineering at Bilkent University, Ankara.

His current research interests include the development of new graph theoretical approaches for medical image analysis, especially for automated cancer diagnosis and grading.

Celal Cigir received the B.S. degree (with high honor) in computer engineering in 2008 from Bilkent University, Ankara, Turkey, in 2008, where he is currently working toward the M.S. degree with the Department of Computer Engineering.

His current research interests include image classification and segmentation, content-based image retrieval, and machine learning.

Cenk Sokmensuer received the Medical degree and pathology training from School of Medicine, Hacettepe University, Ankara, Turkey.

He is currently a Professor of pathology with Hacettepe University. He was a Visiting Scholar with Harvard University during 2003 and 2004, with Necker Children Hospital in France during 1998, and with Victor Dupuy Hospital, France, during 1992. His current research interests include pathology of gastrointestinal system, liver, and endocrine system.

Cigdem Gunduz-Demir (M'05) received the B.S. and M.S. degrees in computer engineering from Bogazici University, Istanbul, Turkey, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from Rensselaer Polytechnic Institute, Troy, NY, in 2005.

In 2006, she joined Bilkent University, Ankara, Turkey, where she is currently an Assistant Professor with the Department of Computer Engineering. Her current research interests include the construction of new bio-computational models and application of computer vision and machine learning algorithms for diagnostic and prognostic purposes in cancer research, in particular, and in medicine, in general.

Dr. Gunduz-Demir was the recipient of the CAREER Award of the National Scientific and Technological Research Council of Turkey. She is a member of the International Society for Computational Biology.