GRAPH WALKS FOR CLASSIFICATION OF HISTOPATHOLOGICAL IMAGES

Gulden Olgun[1]   *Cenk Sokmensuer*[2]   *Cigdem Gunduz-Demir*[1]

[1] Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
[2] Department of Pathology, Hacettepe University Medical School, Ankara, 06100, Turkey

ABSTRACT

This paper reports a new structural approach for automated classification of histopathological tissue images. It has two main contributions: First, unlike previous structural approaches that use a single graph for representing a tissue image, it proposes to obtain a set of subgraphs through graph walking and use these subgraphs in representing the image. Second, it proposes to characterize subgraphs by directly using distribution of their edges, instead of employing conventional global graph features, and use these characterizations in classification. Our experiments on colon tissue images reveal that the proposed structural approach is effective to obtain high accuracies in tissue image classification.

Index Terms— Graphs, graph walks, subgraphs, automated cancer diagnosis, histopathological image analysis.

1. INTRODUCTION

Histopathological examination is a routine practice for cancer diagnosis and grading. This practice includes examining a tissue under a microscope to detect visual changes observed in its appearance as a consequence of cancer. Although it is the gold standard in the current practice of medicine, this examination is time-consuming and may contain a considerable amount of subjectivity. Automated histopathological image classification systems have been proposed to mitigate these problems. These systems model visual changes by extracting mathematical features from tissue images and use the extracted features in their classification. Majority of the studies use one of the two approaches for feature extraction: textural and structural. Textural studies assume that changes occurred as a consequence of cancer can be modeled as texture changes in histopathological tissue images. Thus, they use texture features defined by various techniques including color histograms [1], cooccurrence matrices [2], multiwavelet coefficients [3], local binary patterns [4], and fractal geometry [5]. However, since these techniques define their textures on image pixels, they are susceptible to image variations and noise that are typically observed at the pixel-level.

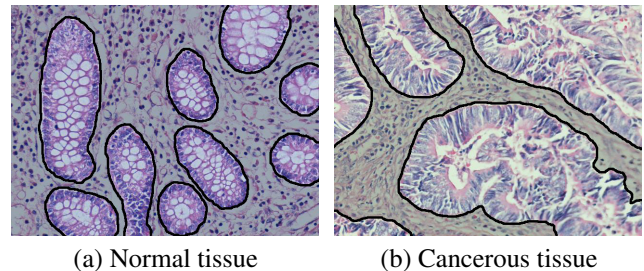|	
(a) Normal tissue	(b) Cancerous tissue

Fig. 1. A colon tissue image contains glandular and nonglandular regions. For colon adenocarcinoma diagnosis, glandular regions provide more important information. This figure shows boundaries of glandular and nonglandular regions in black and nonglandular regions by gray shading.

To alleviate negative effects of pixel-level variations and noise, structural studies define features at the component-level. These studies rely on the fact that visual changes arise from changes occurred in the distribution of histological tissue components. Thus, they model this distribution by representing images with graphs. There exist many studies that construct their graphs on nuclear tissue components [6, 7, 8]. In our recent study, we introduce color graphs constructed on multiple types of tissue components instead of only the nuclear ones [9]. All these studies extract global features (e.g., average degree, diameter, average edge length) from their constructed graphs to quantify tissue images.

Global features defined on entire graph structures are typically aggregates of the properties of graph nodes and edges. Thus, these features may lead to incorrect representations especially when the image that a graph corresponds to is relatively large and contains nonuniform regions in terms of their importance in classification. For example, a colon tissue image contains glandular and nonglandular regions (Fig. 1). Colon adenocarcinomas, which account for 90-95 percent of all colorectal cancers, cause changes in glandular regions. Thus, these regions are more important in the diagnosis of this cancer type. On the other hand, when global features are extracted from the entire graph, nodes/edges located in both glandular and nonglandular regions affect the features in the same way. This may, however, weaken the representation power of the extracted global graph features.
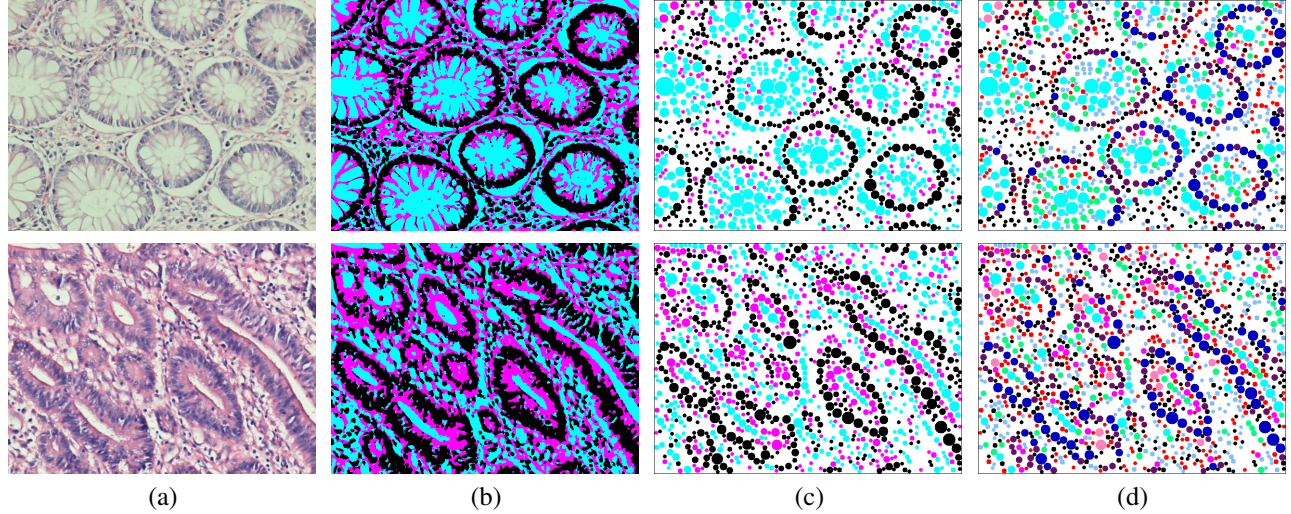
Fig. 2. Graph node definition: (a) original images, (b) quantized image pixels [purple, pink, and white pixels are shown in black, pink, and cyan, respectively], (c) three circle types defined on the quantized pixels, and (d) nine circle types defined based on the previous three circle types and the circles' radii. Here the first row corresponds to a normal tissue and the second one corresponds to a cancerous tissue.

To address this problem, we introduce a new structural approach. In this approach, we propose to represent a tissue image with a set of subgraphs that are obtained by making use of graph walks, characterize the subgraphs with the distribution of their edges, and use these characterizations for classifying the image. The proposed approach differs from the previous studies in the following aspects. First, it represents an image with a set of smaller subgraphs instead of representing it with a single large graph. This subgraph representation gives the potential to extract more distinctive features that locally characterize important regions in an image. Second, it quantifies a subgraph by directly using the distribution of its edges, instead of using conventional global graph features. Working with 3236 microscopic images of colon tissues, the experiments show that our proposed approach leads to high accuracies for all classes in histopathological image classification.

2. METHODOLOGY

The proposed approach obtains multiple subgraphs by graph walking, characterizes them with the distribution of their edges, and uses these characterizations in tissue classification. The details are given in the following subsections.

2.1. Subgraph generation and characterization

We first construct graph G for entire tissue image I and then sample this graph to obtain subgraphs $\mathcal{S} = \{S_1, ..., S_N\}$. We construct graph G by defining a group of circular nodes of different types and assigning edges in between these nodes using Delaunay triangulation.

In this work, circular nodes are defined using the following algorithm. First, we quantize image pixels into three groups that correspond to the main colors (purple, pink, and white) in tissues stained with hematoxylin-and-eosin. For that, we separate the hematoxylin and eosin components of the image using the deconvolution method proposed in [10]. Let h_i and e_i be the hematoxylin and eosin component values of pixel p_i, and h_{avg} and e_{avg} be the average of these values over all pixels. We assign pixel p_i to a group as follows:

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

Then, we separately locate a set of circles on the pixels of each group using the circle-fit algorithm [11]. In this algorithm, we consider only the circles whose radii are greater than 3. After that, we divide the circles of each group into three further subgroups based on their radii, and thus, obtain a total of nine different circle types. Here the first subgroup contains the circles with radii of 4 and 5, the second subgroup contains those with radii of 6 and 7, and the last one contains the remaining circles. At the end of this algorithm, we take these circles as the graph nodes. In Fig. 2, this algorithm is illustrated on two example tissue images.

After defining graph G on entire tissue image I, we obtain subgraphs $\mathcal{S} = \{S_1, ..., S_N\}$ by walking over graph G using the breadth first search (BFS) algorithm, which traverses graph nodes level-by-level starting from an initial node. Particularly, after traversing a node, the BFS algorithm puts all unvisited neighbors of this node into a queue and repeats this procedure for the next node in the queue. In our method,

we start the BFS algorithm from each of the N largest white circles in graph G and obtain a subgraph S_k for each of these starting nodes. In subgraph generation, we terminate the BFS algorithm after it traverses L nodes in the graph. Subsequently, we take the visited L nodes and the edges in between these nodes to generate the subgraph.

Here it is worth to noting that the number L of visited nodes is less than the number of nodes in graph G. Thus, a generated subgraph is expected to cover a smaller local region in tissue image I that graph G represents. Thus, one can consider that entire subgraph set $\mathcal{S}$ corresponds to different local regions in the image. This, in turn, increases the likelihood of representing an image with more distinctive features that focus on its local regions.

After obtaining subgraphs, we characterize each of them by extracting features based on the distribution of its edges. For that, we assign each edge of a subgraph to an edge type based on the node types of its end points. (Here an edge is assigned to one of the 45 types as there are nine node types in the graph.) Then, we quantify the edge distribution based on the edge types. We calculate the frequency of every edge type in the subgraph and use these frequencies as the subgraph's features. In our method, we directly use the distribution frequencies instead of extracting aggregate features (such as average and standard deviation) from the distribution because we do not want to lose any information due to aggregating and because the number of features (which corresponds to the number of edge types) is relatively low. However, it is possible to use different aggregate features or extract the subgraph's global features. Exploring these possibilities could be considered as a future research direction for this work.

2.2. Tissue classification

We classify tissue image I using its subgraphs. To this end, we first classify each subgraph $S_k \in \mathcal{S}$ using its extracted features. Then, we combine the class of every subgraph in a voting scheme to find the class of image I. In this work, we use a linear kernel support vector machine (SVM) classifier for learning. Here we generate three subgraphs from each training image, as explained in the previous subsection, and learn a model on these subgraphs.

3. EXPERIMENTS

3.1. Dataset

The dataset contains 3236 microscopic images of colon tissues stained with hematoxylin and eosin. The images are acquired using a $20\times$ microscope objective lens and 480×640 pixel resolution. We label each of these images with a normal, low-grade cancerous, or high-grade cancerous class. We randomly divide these images into the training and test sets. The training set includes 510 normal, 859 low-grade cancerous, and 275 high-grade cancerous tissues. The test set includes 491 normal, 844 low-grade cancerous, and 257 high-grade cancerous tissues. Note that the training and test sets do not contain any tissue images taken from the same patient.

3.2. Comparisons

We empirically compare our proposed approach (*GraphWalk*) against three previous algorithms. The first one is the *DelaunayTriangulation* algorithm that quantifies an image by constructing a Delaunay triangulation on its nuclear components and extracts features on the constructed graph. The second algorithm is the *ColorGraph* approach that constructs a graph on components of three different types (the three circle types that we also employed in the first phase of node type definition), colors graph edges based on the types of its end nodes, and extracts global features on this color graph [9]. These two algorithms also use linear kernel SVM classifiers. The third algorithm is the resampling-based Markovian model (*RMM*) that also employs a sampling approach [12]. In the *RMM*, a set of sequences are generated on a tissue image. In sequence generation, it selects random points on the image, characterizes these points using their nearby pixels, and orders them based on proximity. Later, it uses Markov models to classify each sequence and then combines the classes of sequences in a voting scheme. For the details of these three algorithms, the reader is referred to our previous work [12].

The *RMM* is similar to the proposed *GraphWalk* approach in the sense that they both use multiple samples representing a tissue image and combine the decisions made on these samples in a voting scheme. On the other hand, the *RMM* uses sequences of points randomly selected in the image whereas the *GraphWalk* algorithm uses subgraphs generated by walking over the graph that represents the image.

To further explore the effectiveness of the proposed approach, we implement two more algorithms. The *BagOfNodes* algorithm uses the distribution of the nine node types in an image. We implement it to understand the importance of constructing a graph instead of directly using the nodes. The *BagOfAllEdges* algorithm considers all graph edges and uses the distribution of their types. This algorithm uses frequency features that are same with those used by our *GraphWalk* approach. However, it extracts these features on the entire single graph whereas the *GraphWalk* approach extracts them on multiple subgraphs. Both of the *BagOfNodes* and *BagOfAllEdges* algorithms use linear kernel SVM classifiers.

3.3. Results

The proposed *GraphWalk* approach has two model parameters: the number N of subgraphs and the number L of visited nodes in each subgraph. The number N determines the size of the subgraph set $\mathcal{S}$, and hence the number of decisions to be voted for classification. Selecting too small values decreases the number of decisions to be voted, which may yield worse

Table 1. Test set accuracies obtained by the proposed *Graph-Walk* approach and the comparison algorithms.

	Normal	Low	High	Avg
GraphWalk	93.48	90.40	92.61	92.17
ColorGraph	92.67	82.46	86.38	87.17
DelaunayTri	89.61	71.56	87.55	82.91
RMM	95.64	87.77	88.56	90.66
BagOfNodes	87.58	89.81	85.60	87.66
BagOfAllEdges	94.50	90.64	89.49	91.54

accuracies. Selecting too large values only slightly affect accuracies, but increases computational time. The number L determines the locality of a generated subgraph. When it is selected too small, only a few nodes are visited and the subgraph covers a very local region. When it is selected too large, too many nodes are visited and the subgraph converges to the entire graph, which results in losing locality. Both of these conditions lower accuracies. In our experiments, we select $N = 50$ and $L = 600$, considering these trade-offs.

In Table 1, we present the test set results obtained by the *GraphWalk* approach and the comparison algorithms. Here we report the accuracies obtained for each class as well as the average accuracy obtained over the three classes. This table shows that the proposed *GraphWalk* approach leads to high accuracies ($> 90\%$) for all of the three classes.

When we compare our approach with the *DelaunayTri-angulation* and *ColorGraph* algorithms, which extract global features on their graphs, we observe that our features extracted on the new graph type are more effective for tissue image classification. The results also show that the direct use of the node type distribution without constructing a graph (the *BagOfNodes* algorithm) yields lower accuracies. Moreover, the comparison between the proposed *GraphWalk* approach and the *RMM* shows that the use of a set of subgraphs (samples) in tissue image representation gives better results than using a set of sequences made up random points. This indicates the effectiveness of graph walking to obtain the multiple samples. Finally, when we compare our approach with the *BagOfAllEdges* algorithm, we observe that the use of multiple smaller subgraphs instead of using a single large graph leads to better accuracies, especially for the high-grade cancerous class. This is attributed to the observation that these smaller subgraphs might better represent local regions in the image.

4. CONCLUSION

This paper presents a new structural approach, in which subgraphs are generated and used for histopathological tissue image representation and classification. This approach obtains a set of subgraphs through graph walking, characterizes these subgraphs using the distribution of their edges, and uses them in classification. The experiments on 3236 colon tissue im-

ages show that the proposed approach is effective to obtain high accuracies for colon adenocarcinoma classification.

As a future work, one could explore using different graph walks to generate subgraphs. As another future work, it is possible to generate and use sequences obtained from these walks instead of constructing subgraphs. In that case, one can consider applying other classifiers defined on sequences.

5. REFERENCES

[1] A. Tabesh et al., "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Trans. Med. Imaging*, vol. 26, no. 10, pp. 1366-1378, 2007.

[2] S. Doyle et al., "A boosted Bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies", *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1205-1218, 2012.

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

[4] O. Sertel et al., "Computer-aided prognosis of neuroblastoma on whole slide images: classification of stromal development," *Pattern Recognit.*, vol. 42, no. 6, pp. 1093-1103, 2009.

[5] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imaging*, vol. 28, no. 7, pp. 1037-1050, 2009.

[6] A. N. Basavanhally et al., "Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 642-653, 2010.

[7] S. Doyle et al., "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," *Proc. Int. Symp. on Biomedical Imaging: From Nano to Macro*, 2008.

[8] C. Demir et al., "Learning the topological properties of brain tumors," *IEEE ACM T. Comput. Bi.*, vol. 2, no. 3, pp. 262-270, 2005.

[9] D. Altunbay et al., "Color graphs for automated cancer diagnosis and grading", *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 665-674, 2010.

[10] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, pp. 291-299, 2001. Source code is available at http://www.dentistry.bham.ac.uk/landinig/software/cdeconv/cdeconv.html.

[11] A.B. Tosun et al., "Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection," *Pattern Recognit.*, vol. 42, no. 6, pp. 1104-1112, 2009.

[12] E. Ozdemir et al.,"A resampling-based Markovian model for automated colon cancer diagnosis", *IEEE Trans. Biomed. Eng.*, vol. 59, no. 1, pp. 281-289, 2012.