Structural And Statistical Pattern Recognition Based Tissue Classification

Athira.v
Student, Dept. of ECE
YCET, Kollam
Kerala
athira4552@gmail.com

Anvesh.P.Thankachan
Associate professor, Dept. of ECE
YCET, Kollam
Kerala
anveshpt2008@gmail.com

Abstract— Cell is the basic functional unit of an organism. The cells together form a tissue and tissues together form an organ. Cancer causes deviations in the distribution of cells, leading to changes in biological structures that they form. Correct localization and characterization of these structures are crucial for accurate cancer diagnosis and grading. A new hybrid model is proposed here that employs both structural and statistical pattern recognition techniques for tissue image classification. This hybrid model relies on representing a tissue image with an attributed graph of its components, defining a set of smaller query graphs for normal gland description, and characterizing the image with the properties of its regions whose attributed sub graphs are most structurally similar to the queried graphs. In order to identify the most similar regions ,the proposed hybrid model searches each query graph over the entire tissue graph using structural pattern recognition techniques and locates the attributed sub graphs whose graph edit distance to the query graph is smallest. It then uses graph edit distances together with textural features extracted from the identified regions to model tissue deformations. The identified regions in cancerous tissues are expected to be less similar to the query graphs than those in normal tissues.

Index Terms— cancer, colon tissue, Graph embedding, Structural pattern recognition, Query graph

I. INTRODUCTION

Every day within our bodies, a massive process of destruction and repair occurs. The human body is comprised of about fifteen trillion cells, and every day billions of cells wear out or are destroyed. In most cases, each time a cell is destroyed the body makes a new cell to replace it, trying to make a cell that is a perfect copy of the cell that was destroyed because the replacement cell must be capable of performing the same function as the destroyed cell. During the complex process of replacing cells, many errors occur. Despite remarkably elegant systems in place to prevent errors , the body still makes tens of thousands of mistakes daily while replacing cells either because of random errors or because there are outside pressures placed on the replacement process that promote errors. Most of these mistakes are corrected by additional elegant systems or the mistake leads to the death of the newly made cell, and another normal new cell is produced. Sometimes a mistake is made, however, and is not corrected. Many of the uncorrected mistakes have little effect on health, but if the mistake allows the newly made cell to divide independent of the checks and balances that control normal cell growth, that cell can begin to multiply in an uncontrolled manner. When this happens a tumor can develop.

Colon adenocarcinoma, which accounts for 90%–95% of all colorectal cancers, originates from epithelial cells and leads to deformations in the morphology and composition of gland structures formed of the epithelial cells (Fig. 1). Moreover, the degree of the deformations in these structures is an indicator of the cancer malignancy (grade).

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

This approach models a tissue image by constructing an attributed graph on its tissue components and describes what a normal gland is by defining a set of smaller query graphs. It searches the query graphs, which correspond to nondeformed normal glands, over the entire tissue graph to locate the attributed sub graphs that are most likely to belong to a normal gland structure. Features are then extracted on these sub graphs to quantify tissue deformations, and hence to classify the tissue. This approach includes three steps: graph generation for tissue images and query glands, localization of key regions that are likely to be a gland, and feature extraction from the key regions.

A. Tissue Graph Representation

In this section, a tissue image is modeled with an attributed graph \( G = \{ V, E, \mu \} \) where \( V \) is the set of nodes, \( E \) is the set of edges, and \( \mu \) is a mapping function. This graph representation relies on locating the tissue components in the image, identifying them as graph nodes and assigning the graph edges between the nodes based on their spatial distribution. However, the exact localization of the components emerges a different segmentation problem an approximation is used that defines circular objects to represent the components.

The image pixels have to be quantized into two groups in order to define these objects: nucleus and non-nucleus pixels. For that the stain is separated using deconvolution method and threshold it with Otsu’s method. Then a set of circular objects is located on each group of pixels using circle fit algorithm. This approximation gives two groups of objects: one group defined on the nucleus pixels and other group defined on the non-nucleus pixels. These groups are called “nucleus” and “white” objects. Note that in this approximation there is not always one-to-one correspondence between the components and objects. For instance a nucleus component typically corresponds to a single nucleus object while a luminal component usually corresponds to many white objects.

After defining the objects as graph nodes their spatial relation is encoded by constructing a tissue graph using Delaunay triangulation.

B. Query Graph Generation

Query graphs are sub graphs that correspond to a normal gland structure in an image. To define a query graph \( G_q \) on the tissue graph \( G \) of a given image, a seed node (object) is selected and it is expanded on the graph \( G \) using the breadth first search (BFS), until a particular depth is reached. Then the visited nodes and edges between nodes are taken to generate the query graph.

The mapping function \( \mu \) attributes each selected node with a label according to its object type and the order in which the node is expanded by the BFS algorithm. Four labels are defined: \( \alpha_{n, in} \) and \( \alpha_{w, in} \) for the nucleus and white objects whose expansion order is less than the BFS depth and \( \alpha_{n, out} \) and \( \alpha_{w, out} \) for the nucleus and white objects whose expansion order is equal to BFS depth. It is worth noting but last two labels are used to differentiate the nodes that form the outer parts of a gland

from the inner ones and this differentiation helps preserve the gland structure better.

![Fig.2. An illustration of generating a query graph](image)

The query graph generation and labeling process are illustrated in Fig.1. In this figure a query graph is generated by taking the dash bordered white object as the seed nodes and selecting the depth as 4. This graph illustration uses a different representation for the nodes of a different label. It uses black circles for \( \alpha_{n, in} \), white circles for \( \alpha_{w, in} \), black circles with green borders for \( \alpha_{n, out} \) and white circles with red borders for \( \alpha_{w, out} \). It also indicates the expansion order of the selected nodes inside their corresponding circles; note that the order is not indicated for the unselected ones.

The search process for key region localization uses the same algorithm to obtain sub graphs to which a query graph is compared. However, these sub graphs are generated by taking each object as the seed node and selecting the depth as the same with that of query graph. Thus the search process for key region localization involves no manual selection.

C. Key Region Localization

This localization of key regions in an image includes a search process. This process compares each query graph \( G_q \) with sub graphs \( G_s \) generated from the tissue graph of the image and locates the ones that are the N-most similar to this query graph. The regions corresponding to the located sub graphs are considered as the key regions. Since a query graph is generated as to represent a normal gland, the located sub graphs are expected to correspond to the regions that have the highest probability of belonging to a normal gland. Typically, the sub graphs located on a normal tissue image are more similar to the query graph than those located on a cancerous tissue image. Thus, the similarity levels of the located sub graphs together with the features extracted from their corresponding key regions are used to classify the tissue image.

The search process requires inexact graph matching between the query graph and the sub graphs, which is known to be an NP-complete problem.
1) Query Graph Search: Let $G_q = \{V_q, E_q, \mu\}$ be a query graph and $v_q \in V_q$ be its seed node from which all the nodes in $V_q$ are expanded using the BFS algorithm until the graph depth $d_q$ is reached. In order to search this query graph over the entire tissue graph $G = \{V, E, \mu\}$, the candidate subgraphs $G = \{V_s, E_s, \mu\}$ are enumerated from the graph $G$. For that, a procedure is followed similar to the one that is used to generate the query graph. Particularly each node is taken that has the same label as a seed node and this node is expanded using the BFS algorithm until the query depth $d_q$. The nodes of the candidate subgraphs are also attributed with the labels in $A = \{a_{in}, a_{out}, a_{in-out}, a_{in-out}\}$ using the mapping function $\mu$, which was used to label the nodes of the query graphs.

After they are obtained, each of the candidate subgraphs $G_s$ is compared with the query graph $G_q$ using the graph edit distance metric and the most similar $N$ non-overlapping sub graphs are selected. To this end the selection is done with the most similar sub graph and eliminate other candidates if their seed node is an element of the selected sub graph. The process is repeated $N$ times until the $N$-most similar sub graphs are selected.

Note that although there may not exist $N$ normal gland structures in an image, the algorithm locates the $N$-most similar sub graphs, some of which may correspond to either more deformed gland structure or false glands. In this study, these glands are not eliminated since the graph edit distance between the query graphs and the sub graphs of more deformed glands are expected to be higher and this will be an important feature to differentiate normal and cancerous tissue images. This feature is especially important in the correct classification of high grade cancerous tissues since sub graphs generated from these tissues are expected to look less similar to a query graph, leading to higher graph edit distances. These higher distances might be effective in defining more distinctive features.

Sometimes there may exist $N$ normal gland structures in an image but the algorithm may incorrectly locate sub graphs that correspond to non-gland tissue regions. However since it locates $N$ sub graphs instead of a single one, the effects of such non-gland sub graphs could be compensated by the others provided that the number of the non-gland sub graphs is not too much. Otherwise, the localization of these sub graphs might lead to misclassifications.

2) Graph Edit Distance Calculation: To select the sub graphs $G_s = \{V_{s}, E_{s}, \mu\}$ that are most similar to the query graph $G_q = \{V_q, E_q, \mu\}$, the proposed model uses the graph edit distance algorithm, which gives error-tolerant graph matching. The graph edit distance quantifies the dissimilarity between a source graph $G_s$ and a target graph $G_q$ by calculating the minimum cost of operations that should be applied on $G_s$ to transform it into $G_q$. This algorithm defines three operations: insertion that inserts a target node into $G_s$, deletion that deletes a source node from $G_s$, and substitution that changes the label of a source node to that of target node. These operations allow matching different sized graphs $G_s$ and $G_q$ with each other.

Fig.3. An illustration of labeling and matching processes used by proposed model: (a) a query graph $G_q$, (b)–(d) two graphs $G_s$ and $G_t$ that are to be matched with $G_q$ and (c)–(e) the matches between the decomposed sub graphs of the query graph $G_q$ and the target graphs $G_s$ and $G_t$. This model uses different attributes for labeling the inner and outer nodes of the same object type. This helps the matching algorithm better optimize the matching cost, preserving the global structure of a gland.

Let $(e_1, \ldots , e_t, \ldots , e_n)$ denotes a sequence of operations $e_i$ that transforms $G_s$ into $G_q$. The graph edit distance $\text{dist}(G_s, G_q)$ is then defined as

$$
\text{dist}(G_s, G_q) = \min \sum_{i=1}^{n} \text{cost}(e_i)
$$

Where cost $(e_i)$ is the cost of the operation $e_i$. Since finding the optimal sequence requires an exponential number of trials within the number of nodes in $G_s$ and $G_q$, the proposed model employs the bipartite graph matching algorithm, which is an approximation to graph edit distance calculation. This algorithm decomposes the graphs $G_s$ and $G_q$ into a set of sub graphs each of which contains a node in the graph and its immediate neighbors.

The bipartite graph matching algorithm can handle matching in relatively larger graphs. However, since it does not consider the spatial relations of the decomposed sub graphs, it may give relatively smaller distance values for the graphs that have different topologies but contain similar sub graphs. This may cause misleading results in the context of our tissue sub graph matching problem. To alleviate the shortcomings of this algorithm, this model proposes to differentiate the nodes that form the outer gland boundaries from the inner nodes by assigning them different attribute labels. For example, suppose we want to match the query graph $G_q$ shown in Fig.3 (a) with two different graphs $G_s$ and $G_t$ that are shown in Fig.3 (b) and (d) respectively. In this figure, considering their object types, the graphs $G_s$ and $G_t$ corresponds to a normal gland whereas the graph $G_q$
corresponds to a non-gland region. The bipartite graph matching algorithm with the definition of inner and outer nodes will compute a larger graph edit distance for the $G_uG_1$ match than the $G_uG_2$ match; the matches found between the decomposed sub graphs of $G_u$ and the target graphs $G_1$ and $G_2$ are illustrated in Fig.3(c) and (e), respectively.

![Graph](image)

III. RESULTS.

Almost all comparison algorithms extract their features on entire tissue pixels of an image. However, this may cause misleading results due to existence of non-gland regions in the image, which are irrelevant to the context of colon adenocarcinoma diagnosis. Moreover, in some images, these irrelevant regions may be larger than gland regions, and hence, they contribute to the extracted features more than the gland regions. As opposed to these previous approaches, this model uses structural information to locate the key regions, which most likely correspond to gland regions, and extract its features from the key regions.

![Graph](image)

Fig. 4. The matches between a query graph $G_u$ and two target graphs $G_1$ and $G_2$ when the model does not differentiate the inner and outer nodes.

D. Feature Extraction and Classification

The tissue image is characterized by extracting two types of local features and classifies the image using neural network classifier. Neural networks are based on approximate models of the brain. The basic building block of an artificial neural network is the neuron. The brain in made up of about 100 billion neurons, with an average of 1,000 to 100,000 input connections per neuron. When many of these neurons are combined they have the properties of a massively parallel super-computer. In multi-layer neural nets, each neuron is connected to other neurons via a weighted communication line.

The weights of the connections are adjusted in training to represent the knowledge of the neural network. One method for adjusting these weights is with a training algorithm. Neural networks are good when dealing with abstract problems, like those based on features and patterns. Artificial neural networks are mainly used in two areas: feature detection and pattern mapping. The local features are extracted by using structural as well as statistical pattern recognition techniques. The first type includes embedding of graph edit distances to the query graph and the second one comprises textural features of the key regions.

![Graph](image)

Fig. 5. Test set accuracies as a function of number $N$ of selected sub graphs.

This hybrid model uses two feature types (structural and textural). It gives better results than the algorithms that combine global textural and structural features, which are defined on the entire image.
IV. CONCLUSION

A hybrid model is introduced that employs both structural and statistical pattern recognition techniques to locate and characterize the biological structures in a tissue image for tissue quantification. The main contribution of the paper is twofold. First, it describes a normal gland in terms of a set of query graphs and models the tissue image by identifying the regions whose sub graphs are most similar to the queried graphs. Second, it embeds the graph edit distances, between the most similar sub graphs and the query graphs and the textural features of the identified regions in a feature vector and uses this vector to classify the tissue image. As opposed to the conventional tissue classification approaches, which quantify tissue deformations by extracting global features from their constructed graphs, the proposed model use graphs directly to quantify the deformations. Additionally, it extracts features on only identified key regions, which allows focusing on the most relevant regions for cancer diagnosis.

This model represents a tissue image as an attributed graph of its components and characterizes the image with the properties of its key regions. The proposed model leads to higher classification accuracies, compared to the conventional approaches that use only statistical techniques for tissue quantification.

REFERENCES


