Colon Biopsy Classification Using Crypt Architecture

Assaf Cohen[1], Ehud Rivlin[2], Ilan Shimshoni[3], and Edmond Sabo[4]

[1] Dept. of Computer Science, Haifa University, Haifa 31905, Israel
[2] Dept. of Computer Science, Technion, Haifa 32000, Israel
[3] Dept. of Information Systems, Haifa University, Haifa 31905, Israel
[4] Dept. of Oncology, Rambam Medical Center and Technion, Haifa 31096, Israel

Abstract. In this paper we introduce a novel method for the detection of diseases in biopsies that contain glandular structures. Most approaches proposed in the literature try to classify the biopsy using the image alone without analyzing its basic elements (such as the nuclei and the glands). The proposed method differs in that it is based on the architecture of the glands in the biopsy and the analysis of each pixel. We demonstrate our novel, three-step method on the task of classifying colon biopsies. First, as described in our previous work, we create a pixel-level classification image, segment the crypts (the glandular structures in the biopsy) using it, and remove false-positive segments. Next, we calculate the crypt architecture using Delaunay triangulation on the crypt centroids and use this architecture to retrieve those crypts that were incorrectly removed in the first step. In the final step, we use the segmented crypts to construct a more accurate architecture and classify each triangle as healthy or cancerous using the classification of the crypts as healthy or cancerous. The method was tested on 54 colon biopsy images: 109 healthy sub-images containing 4944 healthy crypts and 91 cancerous sub-images containing 2236 cancerous crypts. It achieved 92% accuracy in crypt classification and 94% in biopsy region classification.

Keywords: Histology, Colon Crypts, Classification, Architecture

1 Introduction

Colon cancer is the third most common cancer, with nearly 1.4 million new cases in 2012 worldwide [1]. Early detection of cancer can lead to full recovery. Once a possible cancer is detected, it is diagnosed by a biopsy — pathologist's examination of a tissue sample via a microscope. Due to the implication of this diagnosis for the patient, this process is critical. The pathologist needs to be precise and have the ability to sift through huge amounts of data to detect small anomalies in the biopsy. There is thus a clear need for an automatic tool to draw the pathologist's attention to biopsies with suspicious regions.

Considerable progress has been made in the field of histology image analysis and many surveys on biopsy segmentation and classification have been

conducted [2–5]. There are several approaches to this problem. Texture based methods [6, 7] classify the biopsy without understanding its structure and basic elements (such as nuclei and crypts). Relying solely on texture can fail due to changes in color between biopsies due to the staining process and fading of colors over time. Architecture based methods [8, 9] classify the biopsy according to an architecture constructed from the nuclei. This type of architecture is too sensitive to errors in nucleus detection and ignores the crypt features. In [10], a method for classification of prostate glands was proposed using structural and contextual information. We believe that classifying the glands alone cannot produce accurate results.

In this paper a novel biopsy classification method is presented. The problem of classifying the biopsy regions as healthy or cancerous is broken into three major steps. Each step relies on the results of the previous steps but uses a higher level of knowledge to overcome any errors that may have occurred. In our previous work, described in Section 2, the image is Classified at the Pixel Level (denoted by PLC). Using the classified pixels, the crypts are then segmented and false positive crypts identified using a classifier that assigns a certainty score to each crypt. In the second step, described in Section 3, the architecture of the crypts is calculated and used to retrieve missing crypts that were removed in the first step. In the third and final step, described in Section 4, a more accurate architecture is constructed using triangulation on the crypt centroids, including the retrieved crypts. Each triangle is classified as healthy or cancerous using geometry, appearance, the classification of the crypts as healthy or cancerous, and features based on the PLC. The main novelty of this method is the architecture created using the segmented crypts. This type of architecture is general and can be used in the detection of diseases in glandular tissue. Section 5 presents the evaluation of the method and results of each step of the proposed classification method on thousands of healthy and cancerous crypts. We conclude in Section 6.

2 Crypt Segmentation

In this study we work with colon tissue images. The colonic biopsy is composed of a stromal intermedium containing glands (called crypts) and immune system cells that surround the crypts (see Figure 1). We model the crypt as an inner area (lumen, goblet cells, and cytoplasm) with an outer layer of nuclei.

In our previous work [11] (included in the supplementary material), we introduced a novel two-step PLC method and a memory based Active Contour algorithm that we used to segment the colon biopsy crypts. The external forces of the active contour are based on the crypt model and the PLC image.

Due to incorrect selection of crypt candidates or incorrect segmentation, false positive crypts are returned. Features describing the shape of the crypt and the distribution of the pixel classes are extracted from each candidate crypt. A RandomForest (RF) [12] classifier trained on these features eliminates the segments that do not satisfy the crypt model and gives a classification certainty score for each crypt. The false-positive rate drops as a result from 63% to 7%,

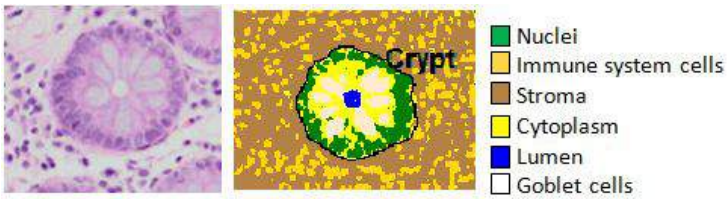

Fig. 1. Healthy colon tissue and crypt structure. A healthy crypt is composed of a lumen and goblet cells in cytoplasm, surrounded by a thin layer of nuclei. The crypt and immune system cells are in the stromal intermedium.

but 15% of the valid crypts are also removed in this phase, whereas prior to this phase, almost all the crypts were retrieved.

3 Missing Crypt Retrieval Using Crypt Architecture

As described in the previous section, the false-positive crypt elimination step uses only features of a single crypt. These features are local and the errors in eliminating valid crypts can be corrected using a higher level of information – the architecture of the crypts.

The architecture is calculated using Delaunay triangulation where the vertices are the centroids of the crypts. The absence of a crypt results in triangles with a different shape from its valid neighboring triangles and the region between the crypts that define the triangle contains eliminated crypts and pixels of crypt classes (see Figure 2a). In order to identify these triangles, a RandomForest classifier is trained based on the following types of features (see Figure 2b):

- **Triangle geometry**. The average, standard deviation, and median of the side lengths, the standard deviation, median of the angles, largest angle, and the area of the triangle.
- **Crypt inter-area**. The percentage of pixels from each crypt class (nuclei, cytoplasm, goblet cells, lumen) inside the area between the crypts. This percentage is determined from the PLC image.
- **Triangle neighborhood geometry**. The number of neighbors of each vertex, standard deviation and average of the distances of a vertex from its neighbors, and the same for the part of the edge in the crypt inter-area.
- **Probability to contain false negative crypts**. The number of invalid crypts, average certainty of the invalid crypts, and the highest crypt certainty inside the triangle.

The classifier is used to find the triangles suspected of containing false negative crypts (see Figure 2a). The architectural information gives a prior that enables us to lower the threshold for the certainty of a valid crypt. Every eliminated crypt that is in a suspicious triangle and whose certainty is above the new empirically selected threshold is considered valid (see Figure 2c). After this step, the false negative rate drops from 15% to 6.3%. Hence the crypt architecture is

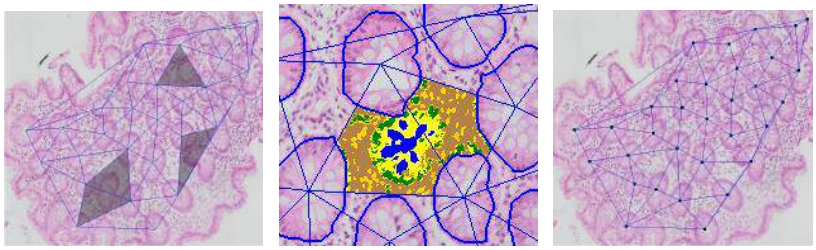

Fig. 2. (a) The architecture of the crypts using Delaunay triangulation on the crypt centroid. In red, the missing crypts. The gray triangles are classified as containing missing crypts. (b) Zoom on a triangle containing a missing crypt. For each triangle, features are extracted on the basis of geometry, content, and similarity to neighboring triangles. (c) The architecture of the crypts after adding those with certainty above the threshold, inside the gray triangles.

more accurate, which is crucial for the partition of the biopsy to cancerous and healthy regions.

4 Biopsy Classification

A biopsy can be healthy or cancerous, but in early stages of cancer development a healthy biopsy might have some cancerous regions (see Figure 6e). In order to accurately segment the biopsy into healthy and cancerous regions, the triangles are classified as healthy or cancerous. This calculated triangle is the smallest unit that has enough information to allow for correct classification. One of the major indications of a cancerous region is that it contains cancerous crypts. In order to use this information in the triangle classification step, we first classify the crypts as healthy or cancerous and then classify the triangles.

4.1 Crypt Classification

In this step, the crypts are classified as cancerous or healthy. A cancerous crypt is larger than a healthy crypt, it does not have a circular shape, its nuclei layer is thicker, and the nuclei pixels are darker. To measure these characteristics, the classifier uses features of the crypt that are extracted from the PLC image:

- **Crypt geometry**. Average radius of the crypt and the distance histogram of the nuclei pixels from the skeleton of the crypt (see Figure 3a).
- **Crypt content**. The percentage of each class of the PLC in the crypt and the thickness of the nucleus layer (see Figure 3b).
- **Crypt appearance**. For each crypt class, the average of the RGB and Lab of the pixels that belong to it (see Figure 3c).

Using the PLC, more accurate features can be extracted: the content and internal structure of the crypt can be measured, as can the thickness of the

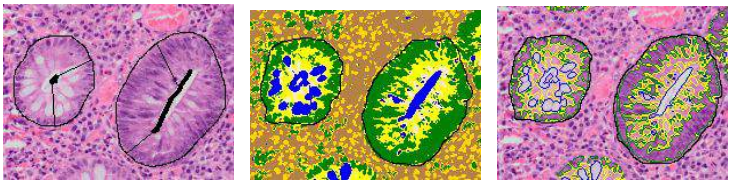

Fig. 3. (a) The radii of the crypt are defined as the distance of the edge pixels from the skeleton. (b) Calculating the percentage of the nucleus, cytoplasm, lumen, and goblet cell pixels. (c) A cancerous crypt has darker nucleus pixels. To measure this characteristic, the average of the RGB pixels is calculated for each of the crypt classes.

nucleus layer, and the average of the RGB channels can be calculated for each crypt class (rather than for the entire crypt).

The RF classifier gives 92% accuracy. Despite its good performance, there are errors in the classification of the crypts. As can be seen in Figure 4, these errors are sparse. They are dealt with during the triangle classification phase by adding the architecture of the crypts and classification of the neighboring crypts.

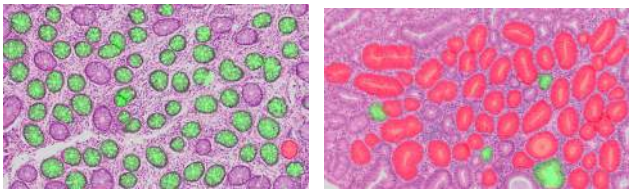

Fig. 4. Results of crypt classification. Left: Healthy biopsy, right: Cancerous biopsy. The green overlay indicates that the crypt is healthy, red indicates that the crypt is cancerous. It can be seen that the classification errors are sparse.

4.2 Biopsy Region Classification

Using the Delaunay triangulation on the crypt centroids, the biopsy is decomposed to small triangular regions that can be used to distinguish cancerous from healthy biopsy regions. A triangle of a cancerous region is larger than that of a healthy region, it differs from neighboring triangles in shape, the crypts that define it are close to each other and different in shape, and the stroma and immune system cell pixels are darker. The classifier uses the following features to measure these characteristics:

- **Triangle geometry and triangle neighborhood geometry**. The same features as in the identification of triangles containing missing crypts, described in Section 3 (see Figure 5a).
- **Crypts' shape similarity**. Variance of the radii of the crypts that define the triangle (see Figure 5a).

- **Crypts' classification**. The classification of the crypts (as described in the previous section) that define the triangle and their classification certainties.
- **Crypt inter-area geometry**. The distance between the edges of the crypts that define the triangle (see Figure 5b).
- **Crypt inter-area content**. The percentage of pixels of each of stroma and immune system cell inside the area between the crypts (see Figure 5b).
- **Crypt inter-area appearance**. The average of RGB and Lab pixels of the stroma and the immune system cells inside the triangle (see Figure 5c).

The RF classifier gives 94% accuracy.

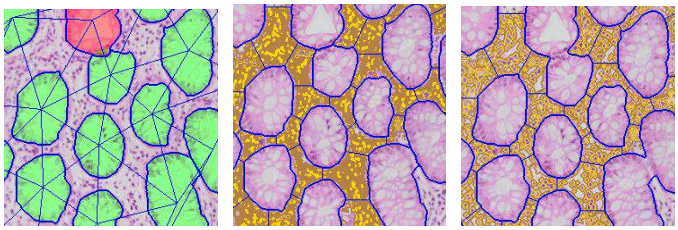

Fig. 5. Triangle Features. (a) Features based on the geometry of the triangles, the similarity of the shape of the crypts that define the triangles, and their classification. (b) Features based on the crypts' intra-area geometry and content. (c) The average of RGB and Lab pixels of the intra-area classes.

5 Results and Discussion

The proposed method is built from three phases. Because each phase relies on the previous phases, in the experiments we evaluate their performance.

Dataset: Since a database with ground truth segmentation of crypts does not yet exist, we collected a database of healthy and cancerous colon biopsies. The biopsies for the database were randomly chosen by E. Sabo MD, a pathologist from the Gyneco-oncology Unit at Rambam Hospital. The database was created by scanning the biopsies under a microscope at x200 magnification. From each scanned biopsy image, sub-images were taken at x4 zoom out. The average size of a sub-image is 800x500 pixels. There were 109 sub-images of healthy colons taken from 33 biopsies and 91 sub-images of cancerous colons taken from 21 biopsies. This database contains 4944 healthy crypts and 2236 cancerous crypts. The ground truth partitioning of the biopsy into cancerous and healthy regions (see Figure 6e) was also confirmed by him.

Missing Crypt Retrieval: The ground truth of the suspicious triangles was created by marking the triangles that contained missing crypts. This step was tested by 5-fold cross-validation on the 200 sub-images and gave 87% accuracy.

The following step is retrieving the crypt segments that have certainty above the threshold of being inside the suspicious triangles. This selective crypt retrieval gives significant improvement in the rate of missed crypts, from 15% to 6.3%, with no loss in segmentation accuracy and an increase in the false-positive rate, from 7.2% to 12.4%.

Crypt Classification: The crypt classification was tested using 5-fold cross-validation. The classifier gives 92% accuracy.

Biopsy Region Classification: The classification of triangles was tested using 5-fold cross-validation using balanced datasets. A triangle is considered as cancerous if it intersects with the red overlay in more than 50% of the triangle area. The classifier gives 94% accuracy, when the classification rates of benign and cancerous triangles are similar, 94.6% and 92.8% respectively. Figure 6 displays the steps of this method. For more results see the supplementary material.

To demonstrate the importance of the type of architecture used we compared Delaunay triangulation generated from nuclei which are lower level features (like the ones used by [8]) to triangulation generated from segmented crypts. Using only color statistics for the pixels in the triangle, the classification accuracy was 71% compared to 83%. This demonstrates that relatively high quality classification can be achieved even with naive features when a meaningful triangulation is used. The method for classifying the triangles is also novel by that it uses new features as input for the classification process. When we replaced the simple color statistics with PLC based features that enable us to analyze the color of the pixels and their distribution from each class, the classification accuracy increased from 83% to 91%. Adding the architecture-based features describing the geometry of each triangle and the similarity to its neighborhood, increased the classification rate to our final result of 94%.

6 Conclusions and Future Work

We have presented a novel method to classify colon crypts and biopsies. The contributions of this paper include: (1) an architecture-based method for the identification of crypts misclassified as invalid; (2) new features for crypt classification using the PLC image; (3) biopsy region classification using the segmented crypts and the PLC image.

In contrast to other architectures defined in previous studies, this one is defined on the basis of crypts. Therefore, the proposed method is general and can be used to classify and estimate the severity of other diseases (such as Crohn's and ulcerative colitis), and can be applied to other types of biopsies that contain glandular structures (such as breast, thyroid, and prostate).

References

1. World Health Organization: World cancer statistics. `http://www.wcrf.org/cancer_statistics/world_cancer_statistics.php` [Online].

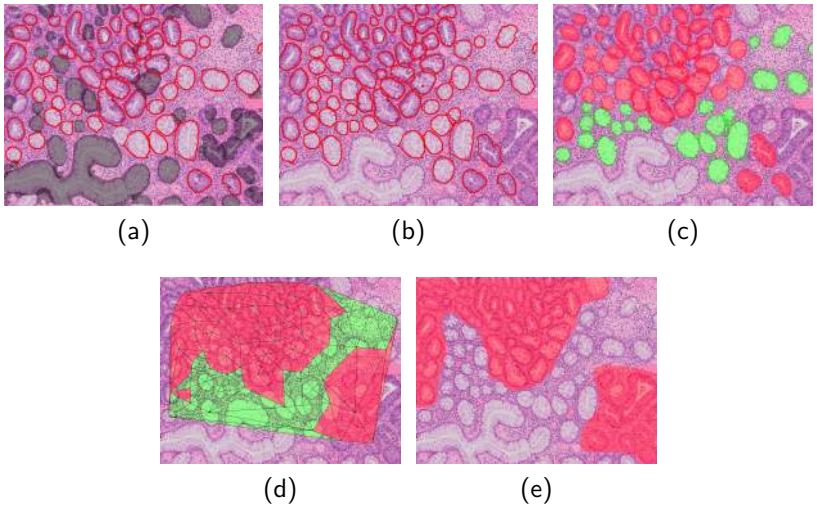

Fig. 6. Example for the steps of the method. (a) Crypt segmentation. The gray overlay indicates that the crypt is classified as false positive. Several valid crypts are misclassified as false positives. (b) Architecture-based retrieval of missing crypts. (c) Crypt classification. (d) Triangle classification. (e) Ground truth image. The red overlay indicates that the region is cancerous.

2. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering **2** (2009) 147–171
3. Demir, C., Yener, B.: Automated cancer diagnosis based on histopathological images: A systematic survey. Rensselaer Polytechnic Institute, Tech. Rep (2005)
4. Belsare, A., Mushrif, M.: Histopathological image analysis using image processing techniques: An overview. Signal & Image Processing **3**(4) (2012)
5. Smochină, C., Herghelegiu, P., Manta, V.: Image processing techniques used in microscopic image segmentation. Technical report, Gheorghe Asachi Technical University of Iaşi (2011)
6. Doyle, S., Feldman, M., Tomaszewski, J., Madabhushi, A.: A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Transactions on Biomedical Engineering **59**(5) (2012) 1205–1218
7. Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K., Zoroofi, R.A.: An image analysis approach for automatic malignancy determination of prostate pathological images. Cytometry Part B: Clinical Cytometry **72**(4) (2007) 227–240
8. Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., Tomaszeweski, J.: Automated grading of prostate cancer using architectural and textural image features. In: ISBI, IEEE (2007) 1284–1287
9. Bilgin, C., Demir, C., Nagi, C., Yener, B.: Cell-graph mining for breast tissue modeling and classification. In: EMBS, IEEE (2007) 5311–5314
10. Nguyen, K., Sarkar, A., Jain, A.K.: Structure and context in prostatic gland segmentation and classification. In: MICCAI. Springer (2012) 115–123
11. Cohen, A., Rivlin, E., Shimshoni, I., Sabo, E.: Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation. Submitted to IEEE Transactions on Medical Imaging
12. Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32