Structural Analysis of Histological Images to Aid Diagnosis of Cervical Cancer

Gisele Helena Barboni Miranda*, Junior Barrera†, Edson Garcia Soares‡ and Joaquim Cezar Felipe*
*Department of Computing and Mathematics, Faculty of Philosophy, Sciences and Languages of Ribeirão Preto, USP
gimiranda@usp.br, jfelipe@ffclrp.usp.br
†Department of Computer Science, Institute of Mathematics and Statistics, USP
jb@ime.usp.br
‡Department of Pathology, Faculty of Medicine of Ribeirão Preto, USP
egsoares@fmrp.usp.br

***Abstract*—The use of computational techniques in the processing of histopathological images allows the study of the structural organization of tissues and their pathological changes. The overall objective of this work includes the proposal, the implementation and the evaluation of a methodology for the analysis of cervical intraepithelial neoplasia (CIN) from histopathological images. For this pourpose, a pipeline of morphological operators were implemented for the segmentation of cell nuclei and the Delaunay Triangulation were used in order to represent the tissue architecture. Also, clustering algorithms and graph morphology were used to automatically obtain the boundary between the histological layers of the epithelial tissue. Similarity criteria and adjacency relations between the triangles of the network were explored. The proposed method was evaluated concerning the detection of the presence of lesions in the tissue as well as the their malignancy grading.**

***Keywords*-Cervical Intraepithelial Neoplasia (CIN); Neighborhood Graphs; Medical Image Processing; Computer-Aided Diagnosis**

I. Introduction

In the last decades, the automatic diagnosis of cancer and the mapping of its evolution have been supported by different methodologies, which can be used to identify primary lesions usually found in the early stages of the disease. It is known that the cure for many types of cancer is associated with early detection and appropriate treatment, according to the malignancy level. Pathologists conduct the assessment of these lesions by the analysis of stained histological sections containing biopsy samples. Generally, the diagnosis is based on international standards. However, this process is still subjective and presents great variability, since the final diagnosis comes from the personal experience of the pathologist [1] [2].

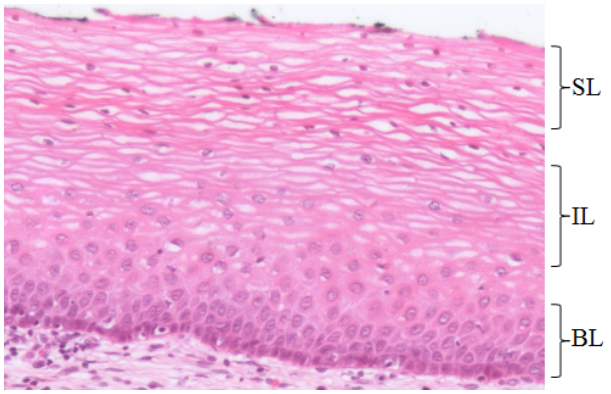
Computer-aided diagnosis in histopathology is based on quantitative measures extracted from intrinsic attributes of images obtained from the histological samples. According to Demir & Yener, the improvement in this research field over the past decades is due to the prospects in the large scale use of decision support systems as part of advanced cancer treatments. Furthermore, it is an area with many challenges to be overcome [3]. Also, improvements in diagnostic accuracy come from enhancements in sample preparation techniques, imaging approaches, as well as training pathologists and other healthcare professionals to better understand and be able to identify key attributes directly associated with the abnormalities that are being detected and evaluated. Thus, decision support systems can be very effective in improving diagnostic capability if they are not constrained by quality of the data that are provided as inputs

In the study reported herein, histopathological images of epithelial tissue of the cervix (Fig. 1(a)) were used as data source to model the structural organization of its cells. The basal layer (BL) of this epithelium presents cells with large nuclei and small rounded-shape cytoplasmic area. The intermediate layer (IL) cells have polygonal shape with vacuoles and glycogen. Finally, the surperficial layer (SL) contains squamous cells normally flat and with no vacuoles [4].

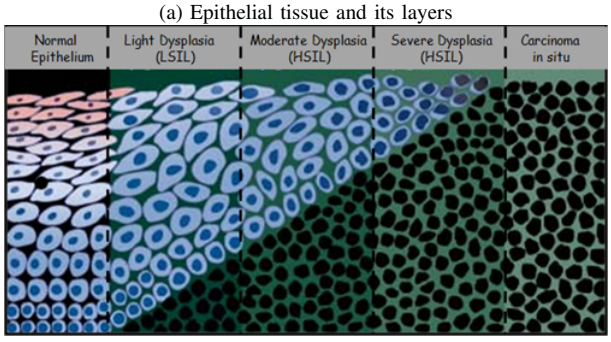
The so-called cervical intraepithelial neoplasia (CINs) consist of proliferative lesions that lead to irregular cell maturation in the tissues. They preceed the squamous cell carcinoma of the cervix, and, if left untreated, they may develop into an invasive carcinoma. CINs can be divided into: mild dysplasia (CIN1), moderate dysplasia (CIN2) and severe dysplasia (CIN3) [4]. Fig. 1(b) shows a schematic representation of the structural changes that occur at the cellular level on the cervix epithelial tissue. From the pathological point of view, it is interesting to note that the CINs vary from mild dysplasias to invasive carcinoma through of a gradual process.

Besides the ratio between the nuclear and the cytoplasmic volume (represented by their areas in two-dimensional images), the portion of the epithelial tissue affected by the lesion is a major perceptual parameter used by the pathologist to define the diagnosis and stands out with high relevance in the analysis: when only the basal layer of the tissue is affected the diagnosis is characterized as low-grade lesion (CIN1), and when the intermediate or superficial layers are affected, the diagnosis is characterized as high-grade lesion (CIN2 or CIN3).

This paper describes the proposal, the implementation and the evaluation of a method for automatic analysis of CINs, through the evaluation of the structural changes caused by these premalignant lesions in the layers of the cervical epithelial tissue. This proposal will be carried out through the study of the processes involved in the structural organization of cells



(a) Epithelial tissue and its layers



(b) Structural variations of the cervical epithelial tissue layers [5]

Fig. 1. Cervix epithelium

by techniques based on Neighbourhood Graphs.

Contributions: The proposed method should contribute to increase the accuracy in generating diagnoses in histopathology, through the automatic identification of the presence of intraepithelial lesions in the tissue and also the identification of their malignancy level. The study and the application of techniques of image processing and pattern recognition appropriate to this context may help in early identification of carcinogenic processes associated with these lesions, which may act as a second opinion to the pathologist.

A. Related work

Landini et al. [6] present an approach based on graph theory for the structural characterization of tissues through mathematical models that describe the geometry of relations between the cells. In this work, images of oral tissues were used to distinguish between cancer dysplastic and normal epithelia.

The use of topological features in the characterization and diagnosis of cancer can also be found in [7]. Here, low-resolution images were used as data source. Metrics were extracted for each cell nuclei representing the nodes of a network. The experiments show an accuracy of at least 85%, which indicates the feasibility of the approach. In a more recent work [8], Gunduz-Demir presents another approach for mapping the evolution of cancer, based on the analysis of connectivity of the network elements.

The characterization of histological tissues through graph theory can also be used to characterize different types of tissues. A methodology for classification of tissue architecture using graph modeling is described in [9]. The proposed methodology was tested in two types of tissues: epithelial and adipose. A similar approach is described in [10], using morphological and topological attributes. Each tissue is represented by a graph obtained from the geometry of its cells, given the proximity between them.

In a more specific context, Keenan et al. [11] presents a method for automatic classification of cervical lesions. The nuclei segmentation in tissue images is based on a process called iterative thresholding. The pixels belonging to nuclei are selected according to specific threshold values at each algorithm iteration. The remaining process generates a network over the segmented nuclei using the Delaunay Triangulation (DT). Then, metrics related to the triangles are calculated. The results indicate better accuracy in the separation of normal and high-level lesions (CIN 3).

B. Technique overview

In the study presented here, a new method for automatic identification of the cervical tissue layers is proposed based on the structural organization of its components. The image acquisition and the identification of the perceptual parameters of interest was performed with the assistance of domain specialists. The watershed transform and a pipeline of morphological operators were used in the segmentation process. The Delaunay Triangulation was the model adopted to represent the cell nuclei as a neighborhood graph. Through the Region-Based analysis the epithelial tissue layers could be identified and used in the classification of the CINs. This method is described in sections III and IV.

II. Technical Background

A. Morphological Reconstruction

Transformations based on morphological reconstructions can be used as filters, eliminating regions of interest. The reconstruction involves the use of a second image M, called "marker", which contains the initial points of the transformation. M is a subset of the input image A. Let A be represented by its connected components, the reconstruction of A by M is denoted $A\Delta_B M$ and is defined through Eq.(1), where B is the structuring element and C_k is a connected component belonging to A:

$$A\Delta_B M = \cup\{C_k : C_k \cap M \neq \emptyset\} \quad (1)$$

The reconstruction can be defined as an important morphological operation with many practical applications, such as the conditional dilation. If an image A is dilated (Eq.(2)) by a structuring element B whose origin is contained in it, A will suffer an expansion which is conditioned by the shape of B. So, the application of successive morphological dilations leads to the loss of the original boundaries of A. This situation can be circumvented by defining the conditions for this expansion, i.e., through the restriction of the translations [12]. Let S be a subset of A, the conditional dilation can be defined as Eq.(3):

$$A \oplus B = \cup_{b \in B}(A_b) \quad (2)$$

$$S \oplus_A B = \cup_{s \in S}(B_s \cap A) \quad (3)$$

The application of n conditional dilations is called geodesic dilation of size n. Considering M a marker, and A the input image, the reconstruction of A by M can be implemented applying a sequence of geodesic dilations until the convergence of the transformation as described in Eq.(4):

$$(M \oplus_A B)^n = (((M \oplus_A B) \oplus_A B) \oplus_A \ldots \oplus_A B) \quad (4)$$

B. Watershed from markers

The watershed transform is one of the most popular methods of segmentation based on region growing. Usually, the watershed is applied on the morphological gradient of the image, however, due to its being susceptible to noise, the resulting image of this operator can present many local minima, generating many watershed lines as final result which is known as over-segmentation. The use of filters on the image of the gradient can reduce this effect [13].

Another approach to avoid over-segmentation is the application of the watershed from markers [14]. The markers bound the "basins" to be segmented, i.e. the minimum points from which the algorithm should start. In this way, each region identified by the algorithm corresponds to a single marker. The watershed from markers can be implemented by reconstruction. This processing is called the minimum imposition.

C. Neighborhood Graphs

The use of models that describe connections between histological components allows the exploration of an additional set of attributes, providing support to the structural analysis of the tissue. The network generation may take into account different criteria to define the links or edges between its components. Assuming a binary image whose connected components are the objects of interest, we have a set V of vertices, represented by these elements and a set E of edges representing neighborhood relations between them. A popular class of models for neighborhood graphs is those obtained from the Voronoi Diagram (VD) [15]. The VD represents a space partition formed by the equidistant points from the elements of V. For all $v \in V$ a polygon $Z(v)$ formed by points closer to v than to any other element of V can be defined in Eq.(5). $Z(v)$ is also called influence zone of v:

$$Z(v) = m \in \Re^2, \forall q \in V \backslash v, dist(m, v) < dist(m, q) \quad (5)$$

Delaunay Triangulation: Also known as the dual graph of the VD, the Delaunay Triangulation (DT) establishes connections among triples of points, always forming triangles. In this model, given a set of points $P = p_1, p_2, , p_n \in \Re^2$, the triangle $p_i, p_j, p_k \in DT(P)$ if its circumcircle is empty (Fig. 2(a)). The duality between the DT and the VD can be seen in Fig. 2(b).

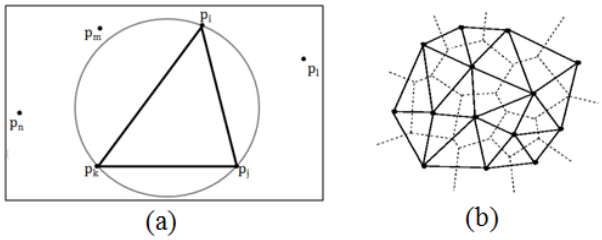


Fig. 2. (a) Criteria for a triangle that belongs to the Delaunay Triangulation (DT). (b) DT (solid lines) vs VD (dotted lines)

III. Structural Analysis

The characteristics of the cell nuclei are important parameters in the analysis of histological images and they can describe specific functional changes. As a consequence, the majority of segmentation methods applied to these images aims to separate the nuclei. The segmentation method applied is determined by the attributes of interest. When nuclei morphology is important, methods based on edge detection are more suitable, as they provide more precise contours. In the case of topological attributes, the approximate location of the nuclei may be sufficient to represent the spatial dependence between them. In this way, the resulting image of the segmentation process is a binary image in which each nucleus is a connected component. Thus, the segmented nuclei can be represented by a set of vertices of a graph using the location of their centroids.

Starting from this set, the relationship among its elements can be modeled through different criteria. Due to its particular properties, the Delaunay Triangulation (DT) was the model adopted in this work. The uniformity of polygons generated by this model (always triangles) allows additional attributes to be explored (Fig. 3). Furthermore, the DT can be easily obtained by the Voronoi Diagram, defined over the segmented nuclei. In this way, a graph $G(V, E)$ can be obtained by the $DT(V)$, where V represents the centroids of the cell nuclei and E is defined by the connections between the elements of V.

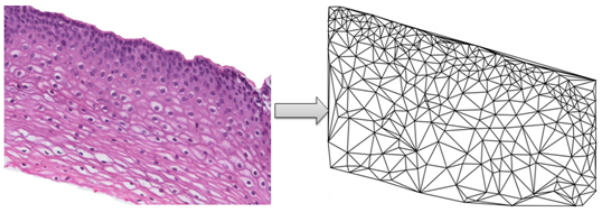


Fig. 3. The Delaunay Triangulation on the tissue

Region-based Analysis: Related studies show that an approach based on the extraction of global attributes (taking into account the entire structure of the tissue) provides support to the analysis of the structural organization of different tissues [9] [10]. However, samples that were obtained from the same tissue can only be differentiated through gradual changes in some tissue regions. As a consequence, for the images analyzed in this work it is more appropriate a local

attribute extraction, characterized by the analysis of tissue regions or clusters. The method proposed in this section aims to define these clusters. Within this context, from the $DT(V)$ the regions of the tissue can be modeled based on adjacency and similarity criterion as described next.

To generate the clusters a new graph $G'(V', E')$ was defined over $DT(V)$, where V', the new set of vertices, is now represented by the set of all the triangles belonging to $DT(V)$ and E' is defined by grouping and adjacency criteria between the pairs of triangles. Two triangles are adjacent if they present at least one vertex in common.

A cluster is defined as a subgraph of G', i.e., the set of vertices V'' of this cluster is a subset of V'. A grouping criterion defines a threshold δ from which a triangle t_i will belong to a subgraph G''. The Euclidean distance was used to define the grouping criterion:

- Triangle Similarity (δ): two triangles t_i e t_j will be in the same cluster if: $dist(t_i, t_j) < \delta$

Where *dist* is the euclidean distance calculated between the vectors formed by the length of the edges of t_i and t_j, and $\delta \in [0, 1]$. The vector generated for each triangle is sorted in ascending order.

Grouping Algorithm: The algorithm described next groups the elements of V'' using the adjacency and the grouping criteria defined above. This algorithm allows the mapping of the DT in clusters that provide a representation of regions of interest in the image under analysis. The grouping criterion (δ) is represented by a percentage of the maximum distance between any two triangles of the network.

The algorithm starts with a reference triangle t_i for which the adjacency and grouping criteria are checked. If $true$, t_j is stacked on P and added to a cluster C_i. As long as P is not empty the same process is repeated analyzing the adjacent triangles to the elements of P, i.e., while there are elements in P, more triangles can be added to the cluster C_i. When the stack is empty the cluster C_i will no longer receive more triangles and a new cluster C_j is created. Then, the process described above is repeated for the remaining triangles until all of them be grouped. Each triangle is visited only once. It is important to note that the checking of adjacency and grouping criteria are always made relatively to the reference triangle t_i. The error rate is a function of δ. By controlling this parameter, it is possible to reduce the variability of the elements in the cluster given the reference triangle.

IV. Automatic Analysis of the Cervix Epithelia

The proposed method aims to identify primary lesions related to cervical cancer. For this purpose, an image database was created, followed by the application of a segmentation algorithm in order to identify the cell nuclei. Then, the Region-Based analysis is applied considering the cell nuclei as nodes. These steps are discussed in the next subsections.

A. Image Acquisition

This work was developed in collaboration with the Cytopathology Laboratory team of the Department of Pathology at Ribeiro Preto School of Medicine, which provided material from cervical uterine histological exams. The digitized microscopic images were acquired from histological sections previously stained with hematoxylin and eosin, containing samples of biopsy exams, using a camera connected to a microscope. The image database was standardized using a 20x objective lens, with an additional increase of 1.6x. The resolution of the digitized images is 1388 x 1040 pixels. The database contains 160 images representing different types of CINs and normal regions.

The images obtained were evaluated with a senior pathologist. This evaluation consisted of the removal of samples that did not present sufficient quality to be classified or not allow a visualization of all layers of the epithelial tissue. Thus, we performed an initial filtering in the data to eliminate possible noises. The expert also determined the CIN classification of each selected image.

B. Segmentation

The process of segmentation aims to separate the cell nuclei in the histological images. For this purpose, a pipeline of morphological operators was applied, followed by the watershed transform (using markers). A specific section of this algorithm concerns the identification of the markers, since they depend on the application context. The next morphological operators were applied in order to obtain the markers for the nuclei:

- *close-by-reconstruction top-hat*: morphological operator defined by the subtraction of the original image by its morphological closing
- *open / closing*: operators applied over a thresholded image.
- *area open*: ensures that the markers will be connected components.

In a second step, the Voronoi Diagram was generated from the internal markers obtained previously. The boundaries defined in this step were used as external markers. Finally, the watershed is applied on the gradient of the original image, using both internal and external markers.

The images were processed in gray scale. This approach leads to a segmentation process less dependent on the acquisition variability of the histological data. Two libraries were used to support the implementation of the segmentation pipeline and the generation of the DT: SDC Morphology Toolbox for C++ and OpenCV (Open Computer Vision) library, also written in C/C++.

C. Identification of the epithelial tissue layers

The structural organization of the epithelial tissue is evaluated over each layer by the pathologist. The layer affected by the lesion defines the malignancy level. The automatic identification of these layers was performed in two steps. In the first one, the boundaries between them were defined by applying the Grouping Algorithm. The adjacency and grouping criteria were evaluated in order to find an optimal δ. For this task, an experiment was performed to adjust this parameter.

In the second step, the clusters were labeled following a supervised approach as described next.

For a normal epithelial image, it is expected to find a large amount of small triangles in the basal layer, triangles of average area in the intermediate layer and, finally, large triangles in the superficial layer. The presence of tissue lesions tends to break this rule, regard the type of triangle. For example, in the presence of CIN3 lesions it is expected to find a large amount of small triangles convering the whole tissue, and, therefore a smaller number of clusters are obtained applying the grouping algorithm.

In this way, three classes of triangles were defined considering their area values: Basal (B), Intermediate (I) and Superficial (S). To obtain the intervals of area values that represent each of these classes, a supervised approach was used: a pathologist was asked to identify manually the boundaries between histological layers in a set of normal images.

After the pathologist segmentation, each image was fragmented into three layers defined according to the boundaries drawn by the pathologist and a DT was generated on each layer. This procedure was performed only once, as a training process. Then, for each image in the training set, three parameters were estimated: the average area of the triangles of the basal layer ($\hat{A}_B$), the average area of the triangles of the intermediate layer ($\hat{A}_I$) and the average area of the triangles of the superficial layer ($\hat{A}_S$). In this way, the classification criterion of a cluster with mean area A_m is:

- *basal, if* $A_m \leq \frac{\hat{A}_B+\hat{A}_I}{2}$
- *intermediate, if* $\frac{\hat{A}_B+\hat{A}_I}{2} \leq A_m \leq \frac{\hat{A}_I+\hat{A}_S}{2}$
- *superficial, if* $A_m \geq \frac{\hat{A}_I+\hat{A}_S}{2}$

The application of the Region-Based analysis, presented in the last section, generates n clusters. After this step, each cluster is classified in one of the labels listed above. This process tends to decrease the number of clusters and can yield one to three clusters. Therefore, the clustering algorithm is important to find similar structures in the network given a grouping criterion (δ) and the labeling of the clusters provides a better representaion of the theoretical model for the CINs grading as described in Figure Fig. 1(b).

D. Feature Extraction

The metrics adopted to compose the feature vectors were chosen based on the structural differences that they provided, such as the occupancy rate (OR) characterized by the sum of the areas of the triangles belonging to a particular layer (A_{C_i}) divided by the sum of the areas of all the triangles in the network (A_C). Also, the mean degree (k_{med}) and the entropy (H) were evaluated.

- *Occupancy Rate*: $OR = A_{C_i}/A_C$
- *Mean Degree*: $k_{med} = 1/N \sum_i k_i$, where k_i is the degree of i-node
- *Entropy*: $H = -\sum_k P(k) log P(k)$, where $P(k)$ is the relative frequency of node degree of value k

E. Classifier

Based on the theoretical model adopted in this work (Fig. 1(b)), a number of possible combinations of clusters were identified. For example, it was found that vectors containing only metrics of basal clusters would be acceptable, since high-grade lesions tend to produce more homogeneous DTs. Differently, the feature vectors obtained from normal images usually present metrics related to the three layers. Finally, for vectors obtained from CIN1 and CIN2 it is expected to find an intermediate number of clusters.

Due to this variation in the number of clusters obtained for each image, four partitions were defined which allowed only vectors of the same size to be compared:

1) B cluster: $\vec{X} = [X_B]$
2) B and I clusters: $\vec{X} = [X_B, X_I]$
3) B, I and S clusters: $\vec{X} = [X_B, X_I, X_S]$
4) any vector containing different metric combinations

The last partition includes, for example, vectors presenting basal and superficial clusters $\vec{X} = [X_B, X_S]$ or intermediate and superficial clusters $\vec{X} = [X_I, X_S]$. These vectors are not representative of the CINs grading as they contain metric combinations that could not represent real situations, therefore they are considered noise and are not classified.

As described in Fig. 4, partitions 1 and 2 lead to CIN3 due to the examples used in the training phase. However, these examples represented only 1.72% of the dataset and the great majority is in partition 3. For this partition, a representative vector ($\vec{X}_{r_y}$) for each class of interest was estimated as follows:

Let Y be the variable that describes the classes of interest and $\hat{E}[X_j]$ the sample mean for the attribute (X_j) considering all the instances associated with class y. For $y \in Y$, a representative vector ($\vec{X}_{r_y}$) is estimated as follows:

$$(\vec{X}_{r_y}) = [\hat{E}[X_1], \hat{E}[X_2], ..., \hat{E}[X_j], ..., \hat{E}[X_A]] \quad (6)$$

Where A is the number of attributes. Then, the euclidean distance is calculated between the representative feature vectors for each class ($\vec{X}_{r_y}$) and the feature vectors of the instances to be labeled (X_i). These representative vectors were obtained by training. The class adopted for a new instance is the class that presents the shortest distance between its representative vector and the vector of (X_i): $d_{min} = d(\vec{X}_{r_y}, \vec{X}_i)$.

V. Results and Discussion

A. Image and Network Segmentation

Fig. 5 shows the results obtained with the application of the pipeline of morphological operators to a normal epithelial tissue image (Fig. 5(a)). Fig. 5(b) shows the same image in gray level and Fig. 5(c) shows the result of the operator close top-hat by reconstruction, highlighting the differences between the nuclei (clusters of darker pixels) and the cytoplasm (clusters of lighter pixels). In Fig. 5(d) is possible to visualize the markers within each nucleus resulting from the application of the following operators: opening, closing and area open. Each

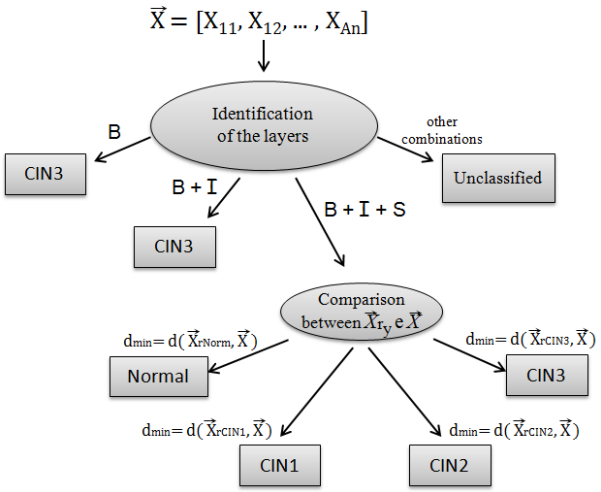
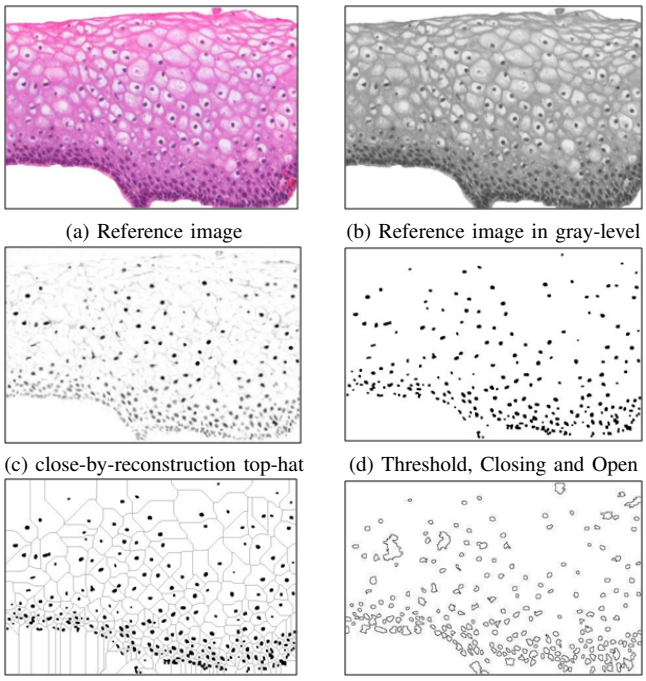


Fig. 4. Classifier

nucleus has only one marker, which is a connected component. This image was thresholded before the application of these operators. Fig. 5(e) shows the Voronoi diagram generated from the internal markers. The boundaries defined in this step characterize the external markers, which were also used in the implementation of the Watershed. Finally, Fig. 5(f) shows the final result of the application of the Watershed, using the Fig. 5(e) as marker.



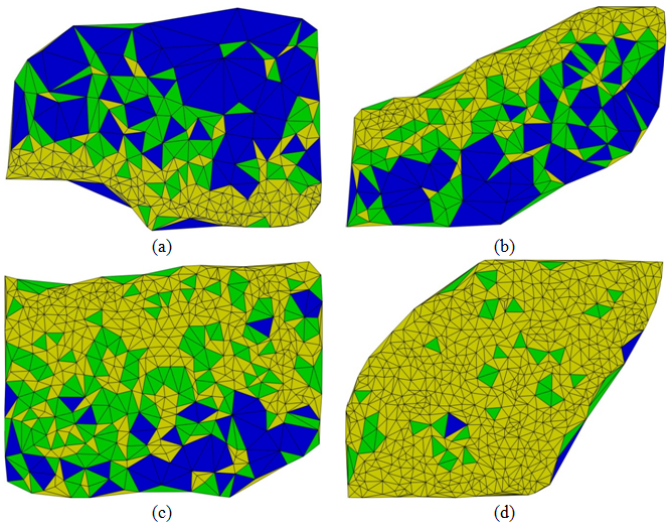
(a) Reference image (b) Reference image in gray-level

(c) close-by-reconstruction top-hat (d) Threshold, Closing and Open

(e) Internal and external markers (f) Watershed

Fig. 5. Steps of the segmentation process

Fig. 6 shows the result obtained by applying the Region-Based analysis over the segmented image. In this figure is possible to analyze the grading of the CINs regarding the network changes. The yellow clusters represent the basal layers, the green ones represent the intermediate layers and the blue clusters represent the superficial layers.



(a) (b) (c) (d)

Fig. 6. Resulting clusters for (a) Normal, (b) CIN1, (c) CIN2 and (d) CIN3 images: basal (yellow), intermediate (green) and superficial (blue)

B. Characterization of the CINs

The experiments were carried out considering each metric alone and, also, different combinations of metrics, in order to evaluate the classification accuracy of CINs for different sets of attributes.

Normal vs Lesion: The proposed method was evaluated concerning the detection of lesions in the tissue. For this purpose, the CINs were grouped in the same class. In this experiment, 60 images were used for training: 30 representing normal cases and 30 representing at least one kind of CIN. Finally, 15 images (10 representing normal images and 5 representing the presence of CINs) were used for test.

To evaluate the metrics, different feature vectors were created by extracting measures of the tissue layers. Sets of attributes were defined by combining the following metrics: OR, k_{med} and H. These metrics were applied to each cluster: basal (B), intermediate (I) and superficial (S). The sets representing the combination metric/cluster are described next:

a) = $[OR\_B, OR\_S]$
b) = $[k_{med}\_B, k_{med}\_I, k_{med}\_S]$
c) = $[k_{med}\_B, k_{med}\_S]$
d) = $[H\_B, H\_I, H\_S]$
e) = $[H\_B, H\_S]$
f) = $[OR\_B, OR\_S, k_{med}\_B, k_{med}\_S]$
g) = $[OR\_B, OR\_S, H\_B, H\_S]$
h) = $[k_{med}\_B, k_{med}\_S, H\_B, H\_S]$
i) = $[OR\_B, OR\_S, k_{med}\_B, k_{med}\_S, H\_B, H\_S]$

For each vector combination, a 5-fold cross-validation was performed alternating images from training and testing. The values presented in Table I are average values of accuracy after cross-validation. The "attributes" column corresponds to the sets of combinations shown above. In this context, accuracy

represents the number of correct classified instances. Besides the mean accuracy (AC_m), Table I also shows the values of the average sensitivity (SS_m) and the average specificity (SP_m) and their respective standard deviations (sd). The values in bold represent the highest values of accuracy, sensitivity and specificity. The best results in the task of detecting abnormalities were obtained using the first set of attributes: (a). For this set, the values of mean accuracy, sensitivity and specificity were 88%, 98% and 68% respectively.

We can observe that the same values of accuracy was obtained for sets (b) and (c) that evaluate the average degree. It shows that the use of measures of the intermediate layer did not alter the accuracy obtained for this data set. A similar analysis can be performed to SS_m and SP_m, especially for the entropy (d and e) and the occupancy rate (a).

TABLE I
EVALUATION OF THE ACCURACY IN DETECTING THE PRESENCE OF LESIONS.

$\vec{X}$	AC$_m$	sd(AC$_m$)	SS$_m$	sd(SS$_m$)	SP$_m$	sd(SP$_m$)
$\vec{X}_a$	**0,88**	0,09	0,98	0,04	**0,68**	0,3
$\vec{X}_b$	0,85	0,09	0,96	0,09	0,64	0,33
$\vec{X}_c$	0,85	0,09	0,96	0,09	0,64	0,33
$\vec{X}_d$	0,83	0,11	**1,00**	0,00	0,48	0,33
$\vec{X}_e$	0,85	0,13	**1,00**	0,00	0,56	0,38
$\vec{X}_f$	0,85	0,09	0,96	0,09	0,64	0,33
$\vec{X}_g$	0,84	0,13	**1,00**	0,00	0,52	0,39
$\vec{X}_h$	0,87	0,09	0,98	0,04	0,64	0,33
$\vec{X}_i$	0,87	0,09	0,98	0,04	0,64	0,33

Normal vs CIN1: The classes analyzed in this experiment show very similar visual patterns, since CIN1 is characterized by structural changes only in the basal layer of the epithelium. The highest accuracy in the comparison between these two classes was obtained also using the first set or attributes (vector X_a) as shown in Table II. The structural changes of CIN1 increase the number of connections between the cells and therefore the mean degree (k_{med}) of the basal layer. When analyzing the sensitivity rates, the entropy (H) provided better identification of true positives. However, the vectors that provided sensitivity rates of 1.0, also provided the lowest rates of specificity.

Similar to the previous experiment, similar values of accuracy were found for the sets of attributes analyzed, due to the high values of standard deviation. In this experiment, 60 images were used during the training, being 30 of each class, and 9 images for test.

CIN1 vs CIN2: The ORs of basal and superficial layers also provided the highest accuracy values when comparing CIN1 and CIN2 (Table III), as well as the average sensitivity and specificity. These classes differ by the tissue layer affected by the lesion which explains why the attribute OR have been highlighted in this experiment for all combination of metrics tested. Also, the ORs presented the best relation between sensitivity and specificity. The analysis of Table III also shows that the use of the entropy of the intermediate layer enhances the absolute accuracy when comparing the vectors X_d and X_e.

TABLE II
EVALUATION OF ACCURACY IN DISTINGUISHING BETWEEN NORMAL AND CIN1 IMAGES.

$\vec{X}$	AC$_m$	sd(AC$_m$)	SS$_m$	sd(SS$_m$)	SP$_m$	sd(SP$_m$)
$\vec{X}_a$	**0,73**	0,23	0,90	0,14	**0,60**	0,47
$\vec{X}_b$	0,67	0,27	0,80	0,27	0,56	0,46
$\vec{X}_c$	0,67	0,27	0,80	0,27	0,56	0,46
$\vec{X}_d$	0,65	0,18	0,95	0,11	0,40	0,40
$\vec{X}_e$	0,67	0,22	**1,00**	0,00	0,40	0,40
$\vec{X}_f$	0,67	0,27	0,80	0,27	0,56	0,46
$\vec{X}_g$	0,67	0,22	**1,00**	0,00	0,40	0,40
$\vec{X}_h$	0,69	0,25	0,90	0,14	0,52	0,46
$\vec{X}_i$	0,69	0,25	0,90	0,14	0,52	0,46

The best rates of sensitivity were also obtained using measures related to the entropy of the three layers. In this experiment, 44 images were used for training being 22 of each class and 15 images were used for test.

TABLE III
EVALUATION OF ACCURACY IN DISTINGUISHING BETWEEN CIN1 AND CIN2 IMAGES.

$\vec{X}$	AC$_m$	sd(AC$_m$)	SS$_m$	sd(SS$_m$)	SP$_m$	sd(SP$_m$)
$\vec{X}_a$	**0,77**	0,14	**0,80**	0,18	**0,77**	0,14
$\vec{X}_b$	0,65	0,06	0,60	0,28	0,67	0,06
$\vec{X}_c$	0,64	0,08	0,60	0,28	0,65	0,07
$\vec{X}_d$	0,65	0,08	0,73	0,13	0,63	0,10
$\vec{X}_e$	0,60	0,11	**0,80**	0,18	0,55	0,11
$\vec{X}_f$	0,65	0,09	0,60	0,28	0,67	0,08
$\vec{X}_g$	0,61	0,09	**0,80**	0,18	0,57	0,09
$\vec{X}_h$	0,67	0,08	0,60	0,28	0,68	0,07
$\vec{X}_i$	0,67	0,08	0,60	0,28	0,68	0,07

CIN2 vs CIN3: The differences between CIN2 and CIN3 can be identified by the layers affected by the lesion. The OR also excelled in this experiment which were carried out to compare these two classes, as outlined in Table IV. Furthermore, the OR also presented the best values of specificity. In this experiment, 40 images were used for training being 20 of each class and 7 images were used for test.

TABLE IV
EVALUATION OF ACCURACY IN DISTINGUISHING BETWEEN CIN2 AND CIN3 IMAGES.

$\vec{X}$	AC$_m$	sd(AC$_m$)	SS$_m$	sd(SS$_m$)	SP$_m$	sd(SP$_m$)
$\vec{X}_a$	**0,86**	0,20	0,70	0,45	**0,92**	0,20
$\vec{X}_b$	0,80	0,31	0,60	0,55	0,88	0,27
$\vec{X}_c$	0,83	0,26	0,70	0,45	0,88	0,27
$\vec{X}_d$	0,77	0,22	0,60	0,42	0,84	0,26
$\vec{X}_e$	0,83	0,23	**0,80**	0,27	0,84	0,26
$\vec{X}_e$	0,83	0,26	0,70	0,45	0,88	0,27
$\vec{X}_f$	0,80	0,24	0,70	0,45	0,84	0,26
$\vec{X}_h$	0,83	0,26	0,70	0,45	0,88	0,27
$\vec{X}_i$	0,83	0,26	0,70	0,45	0,88	0,27

C. Evaluation of the identification of the CINs

In the experiments presented in the previous section, the OR attribute allowed a good separation between the analyzed

classes, providing accuracy values always greater than 73%. This result can be compared to other sets of attributes, showing better results in some cases. However, in general, in all the experiments, the dp values indicate that all sets of attributes tested provided very similar results regarding the values of accuracy, sensitivity and specificity.

The analysis of the values of sensitivity showed that the entropy (H) presented the best results, however, H also presented a very low specificity. Although it is interesting to obtain high rates of sensitivity to the problem under consideration, a low specificity may lead to more aggressive diagnostic conducts in cases that could be applied simpler treatments. For example, a false positive exemplified in Table III could lead to a surgical intervention in a case that would require only a non-intrusive treatment.

VI. Conclusion

This paper has presented a method for automatic analysis of cervical histological images based on the analysis of topological features in order to identify the presence of CINs. This method relies on the characterization of cell clusters or layers with similar characteristics supporting the feature extraction by regions.

This work differs from the work described in Keenan et al [11] in three aspects: the process of identification of the epithelial tissue layers; the use of the properties of the DT; and the classification of the CINs. In the work of Keenan, the epithelial tissue is indistinctly divided into three equal parts and the average area of the triangles presented in the network was used to classify the CINs. In this paper, a new method for automatic identification of the tissue layers was proposed based on the structural organization of its components. In addition, the presented approach is independent of image scale and angular position of histological structures. Furthermore, the area of the triangles of DT was used to identify the clusters and not to classify the CINs. For this task, a specific classifier was designed.

As seen in the results, the relationship between the different types of CINs and the different layers was well represented by the OR metric, showing the gradual transition of the CINs. Accuracy values higher than 70% were obtained when comparing the following classes: Normal x CIN1, CIN1 x CIN2, and, CIN2 x CIN3. When comparing the four classes (Normal x CIN1 x CIN2 x CIN3), the maximun accuracy obtained was 64%. The method described by Keenan et al. provided an accuracy rate of 62% when comparing the three CINs (CIN1 x CIN2 x CIN3). The work of Landini et al. also evaluated the accuracy in classification of premalignant lesions (in this case, related to oral carcinoma), reaching a maximum of 52%, using only samples of high and low degree of malignancy.

Although the presented method was specifically applied in the automatic detection of CINs, we consider it could also be applied in other problems involving structural analysis of histological tissue, covering a more general family of applications. This task is part of future work as well as the evaluation of the clinical applicability of the method.

Acknowledgment

This research was supported by the São Paulo State Research Foundation (FAPESP): grant 2009/04752-1.

References

[1] S. Ismail, A. Colclough, J. Dinnen, D. Eakins, D. Evans, E. Gradwell, J. O'Sullivan, J. Summerell, and R. Newcombe, "Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia." *British Medical Journal*, vol. 298, no. 6675, pp. 707–710, 1989.

[2] W. McCluggage, M. Walsh, C. Thornton, P. Hamilton, A. Date, L. Caughley, and H. Bharucha, "Inter-and intra-observer variation in the histopathological reporting of cervical squamous in traepithelial lesion susing a modified bethesda grading system," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 105, no. 2, pp. 206–210, 1998.

[3] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: a systematic survey," *Rensselaer Polytechnic Institute, Tech. Rep*, 2005.

[4] S. Araujo, *Citologia e Histopatologia Bsicas do Colo Uterino para Ginecologistas / Basic Cytology and Histopathology of the Cervix for Gynecologists*. VP, 1999.

[5] A. Stevens and J. Lowe, *Human Histology*, 2nd ed. Times Mirror International Publishers Ltd, 1997.

[6] G. Landini and I. Othman, "Architectural analysis of oral cancer, dysplastic, and normal epithelia," *Cytometry Part A*, vol. 61, no. 1, pp. 45–55, 2004.

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

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

[9] R. Rizzio, B. Stransky, F. Zampirolli, and J. Barrera, "Classifying biological tissue architectures with small samples," *Unpublished*.

[10] F. de Assis Zampirolli, B. Stransky, A. Lorena, and F. de Melo Paulon, "Segmentation and classification of histological images-application of graph analysis and machine learning methods," in *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*. IEEE, 2010, pp. 331–338.

[11] S. Keenan, J. Diamond, W. Glenn McCluggage, H. Bharucha, D. Thompson, P. Bartels, and P. Hamilton, "An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (cin)," *The Journal of pathology*, vol. 192, no. 3, pp. 351–362, 2000.

[12] E. Dougherty and R. Lotufo, *Hands-on morphological image processing*. Society of Photo Optical, 2003, vol. 59.

[13] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[14] F. Meyer and S. Beucher, "Morphological segmentation," *Journal of visual communication and image representation*, vol. 1, no. 1, pp. 21–46, 1990.

[15] L. Vincent, "Graphs and mathematical morphology," *Signal Processing*, vol. 16, no. 4, pp. 365–388, 1989.