Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering

Marina E. Plissiti, Christophoros Nikou, *Member, IEEE*, and Antonia Charchanti

***Abstract*—In this paper, we present a fully automated method for cell nuclei detection in Pap smear images. The locations of the candidate nuclei centroids in the image are detected with morphological analysis and they are refined in a second step, which incorporates *a priori* knowledge about the circumference of each nucleus. The elimination of the undesirable artifacts is achieved in two steps: the application of a distance-dependent rule on the resulted centroids; and the application of classification algorithms. In our method, we have examined the performance of an unsupervised (fuzzy C-means) and a supervised (support vector machines) classification technique. In both classification techniques, the effect of the refinement step improves the performance of the clustering algorithm. The proposed method was evaluated using 38 cytological images of conventional Pap smears containing 5617 recognized squamous epithelial cells. The results are very promising, even in the case of images with high degree of cell overlapping.**

***Index Terms*—Cell nuclei detection, fuzzy C-means (FCM), morphological reconstruction, Pap smear images, support vector machines (SVMs).**

I. Introduction

THE CORRECT interpretation of the microscopic examination of cells and tissues is crucial for the final diagnostic decision for many diseases. One of the most interesting application fields of microscopic screening is the detection of precursors of cancer in cell samples. Nowadays, the most eminent example is screening for cervical cancer in its early stages, through the well-known Pap smear [1].

The visual interpretation of Pap smear images is a tedious, time consuming, and in many cases an error-prone procedure. This is a consequence of the fact that the conventional smear exhibits uneven layering, crowding, and overlapping of cells. Furthermore, there exist variances in illumination and dye concentration of the cells due to the staining procedure. Also, there are numerous variables, such as air drying, excessive blood, mucus, bacteria, or inflammation, which make the recognition of the suspicious cells a difficult task.

Manuscript received March 22, 2010; revised July 7, 2010; accepted September 22, 2010. Date of publication October 14, 2010; date of current version March 4, 2011.

M. E. Plissiti and C. Nikou are with the Department of Computer Science, University of Ioannina, Ioannina 45110, Greece (e-mail: marina@cs.uoi.gr; cnikou@cs.uoi.gr).

A. Charchanti is with the Department of Anatomy-Histology and Embryology, Medical School, University of Ioannina, Ioannina 45110, Greece (e-mail: acharcha@cc.uoi.gr).

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

Digital Object Identifier 10.1109/TITB.2010.2087030

The large number of cells and the variation in cell types each Pap smear image includes are also factors of complexity. There are generally three types of squamous cells seen on Pap smear images: 1) the superficial cells are the largest of the three and have small pyknotic nuclei and cytoplasm that generally stains eosinophilic (red); 2) the intermediate squamous cells, which are similar in appearance but are slightly smaller in size and have larger, clearly structured, round nuclei with cytoplasm that usually stains basophilic (blue); and 3) the parabasal cell type that is smaller, more rounded, and immature cell type.

The prerequisite for any further processing of these images is the automated detection of cell nuclei, which presents significant changes when the cell is affected by a disease. In pathological situations, the nucleus may exhibit disproportionate enlargement, irregularity in form and outline, hyperchromasia, or irregular chromatin condensation. The identification and quantification of these changes in the nucleus morphology and density contribute in the discrimination of normal and abnormal cells.

The first attempts to detect and segment cells in cervical microscopic images were based on image-thresholding techniques [2]. In addition, pixel classification was also proposed for the segmentation of cervical images [3]. Another class of methods concerns morphological watersheds for the separation of the cytoplasm and the nucleus of each cell [4], [5]. The boundaries of the structuring elements of the cells can be obtained by employing methods based on active contours [6], template fitting [7], [8], genetic algorithms [9], region growing with moving K-means [10], and edge detectors [11], [12].

In Table I, the methods that have appeared in the literature for the segmentation of Pap smear images are presented. As it can be observed, many methods do not take advantage of the color information of the cervical images by converting the color image to its gray-scale counterpart [4], [6]–[12], and therefore, missing the color information. Also, the problem of overlapping cells is not considered in many methods, which identify the borders of the nucleus and the cytoplasm in cervical images that contain only one cell or isolated cells [4], [6], [9], [11], [12].

Considering the general methods that these approaches are based on, we can conclude that the powerful techniques that the mathematical morphology provides for the image segmentation are not efficiently exploited. Even in the case, where morphological watersheds are used in [4] and [5], these methods seem to suffer from several limitations. The method proposed by Bamford and Lovell [4] was applied in gray-scale images of low resolution and results in the identification of the location

TABLE I
ADVANTAGES AND LIMITATIONS OF STATE OF THE ART METHODS FOR PAP SMEAR CELL NUCLEI DETERMINATION

METHOD	YEAR	ADVANTAGES	LIMITATIONS
Bamford. *et al.* [4]	1996	• Simple segmentation method for the determination of the boundaries of the cells. • Ensures closed boundaries.	• Does not handle overlapped cells. • Grayscale images. • Lack of identification of the nucleus boundary.
Bamford.*et al.* [6]	1998	• Ensures closed boundaries for the nucleus and cytoplasm of isolated cells. • High rate of accurate segmentation. • Large number of test images.	• Does not handle overlapped cells. • Grayscale images. • Two captures of the cell image were used.
Wu *et al.* [7] (1)	1998	• Incorporates *a-priori* knowledge about the shape of the cell. • Investigates the case of overlapping breast cells.	• Grayscale images. • Many parameters to be tuned.
Garrido. *et al.* [8]	2000	• A reformulated Hough transform is introduced. • A deformable template model is used for the refinement of the cells boundary.	• Grayscale images. • The method is affected by the excess of edge points or overlapped objects in complex images.
Lezoray *et al.* [5] (2)	2002	• Incorporates color information on the watershed segmentation. • High rate of accurate segmentation.	• A training set is needed for the achievement of best results.
Lassouaoui *et al.* [9]	2003	• Introduces an optimization step based on genetic algorithms to increase the segmentation quality.	• Does not handle overlapped cells. • Grayscale images. • Grayscale images.
Bak *et al.* [3]	2004	• A new criterion function based on statistical structure of the object is used.	• Grayscale images.
Mat Isa. *et al.* [10]	2005	• Region growing based technique in which the seed points locations and the threshold values are determined automatically.	• Grayscale images.
Yang-Mao *et al.* 12]	2008	• A new edge enhancement nuclei and cytoplast contour detector is used. • A new error measurement method is introduced.	• Does not handle overlapped cells. • Grayscale images.
Lin *et al.* [11]	2009	• Ensures closed boundaries.	• Does not handle overlapped cells. • Grayscale images.

(1) The cervical specimen that was used for the acquisition of the test image was stained by the Crocker and Nar staining technique.

(2) The method was applied on images from serous cytology stained with the Pap technique.

of isolated cells in each image. However, cell nuclei that are in cell clusters are not detected. Furthermore, the method proposed by Lezoray and Cardot [5] is based on pixel-classification techniques for the detection of the nuclei markers, in order to avoid the oversegmentation that the watershed algorithm may produce. In pixel-classification techniques, the choice of the number of the classes the pixels belong to plays a crucial role for the final segmentation result. Pap smear images exhibit great complexity and the number of pixel classes is not obvious. The rough assumption that all the pixels of the image are distributed into two classes, such as nuclei pixels and other pixels, would produce noisy results.

In this paper, we propose a novel method for the automated detection of nuclei locations in conventional Pap-stained cervical cell images, which may contain both isolated cells and cell clusters. The method exploits the particular nuclei characteristics through morphological image analysis. In general, the cell nucleus is darker than the surrounding cytoplasm [see Fig. 1(a)]. However, its image intensity value exhibits extensive variation due to the staining procedure or the type of the cell, and sometimes it may coincide with other areas of the image with cell overlapping [see Fig. 1(b)]. If we consider the mapping of the image in the 3-D space [see Fig. 1(c)], we can see that the locations of the nuclei are depicted as intensity valleys. Nevertheless, not all the intensity valleys of the same depth correspond to the location of a nucleus. As we can see in Fig. 1(c), the points A, B, and C belong to different intensity valleys, which approximately have the same depth. However, only the point A belongs to the location of a true nucleus. For the determination of the true nucleus location, the local depth of the intensity valley must be compared with the corresponding local depth of its surrounding area. This figure depicts clearly that the local depth h_A of the point A has higher value than the local depths h_B and h_C of the points B and C, respectively. Based on this fact, we propose an effective method that can distinguish the nuclei locations in Pap smear images.

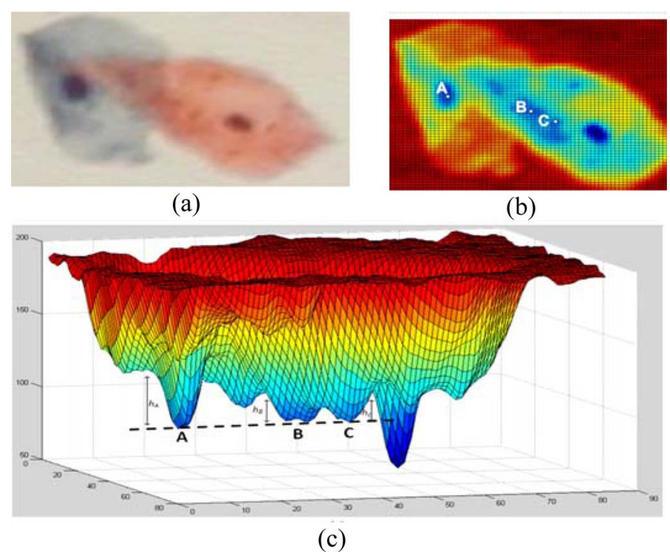


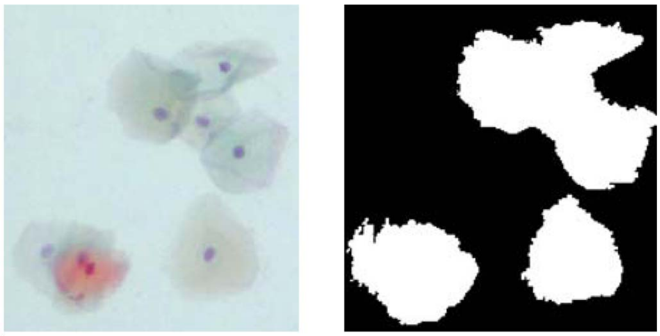
Fig. 1. (a) Initial cell image. (b) Mapping of the intensity values in the color space, where high-intensity values are represented by red and small-intensity values are represented by blue. Point A corresponds to the location of a true nucleus, and points B and C correspond to areas of cell overlapping. (c) Mapping of the initial image in 3-D space. The points A, B, and C are lying in the same intensity level but only point A corresponds to the location of a true nucleus. As it is observed, the local depth h_A of this point is more pronounced with respect to h_B and h_C.

Our paper, whose shorter and preliminary version was presented in [13], consists of four phases: 1) the preprocessing; 2) the detection of candidate cell nuclei centroids; 3) the refinement of candidate cell nuclei centroids; and 4) the decision phase that includes the determination of the final nuclei locations. These phases are described in detail in the following paragraphs.

II. METHOD

A. Preprocessing

The preprocessing phase is necessary for the extraction of the background in order to reduce the searching area in the image. In



(a) (b)

Fig. 2. (a) Initial Pap smear image, and (b) binary mask, which is obtained after the preprocessing step.

the first step, for contrast enhancement and edge sharpening, the contrast-limited adaptive histogram equalization [14] is applied individually to each color component. Next, from each filtered image, a binary image is produced through global thresholding using the method proposed by Otsu [15]. Finally, in the third step, the binary mask BW, with the regions of interest of the image included, is given by

$$BW = BW_1 \cup BW_2 \cup BW_3 \tag{1}$$

where BW_1, BW_2, and BW_3 are the binary masks in the red, green, and blue channels of the initial image. A morphological dilation is then performed in order to expand the boundaries of the region of interest, i.e.,

$$BW = BW \oplus X \tag{2}$$

where X is a 3×3 flat structuring element. After this operation, the connected components with an area smaller than the area of an isolated cell are undesired. For this reason, we remove all connected components with an area smaller than 500 pixels, which is a value smaller than the area of an isolated cell (which in general varies between 900–7000 pixels, determined empirically after careful examination by a cytopathologist) and larger than the size of the small objects. The resulted binary image (see Fig. 2) is used as a mask to indicate the regions, where the detection algorithm is then applied.

B. Detection of Candidate Cell Nuclei Centroids

The areas of interest in the image obtained in the preprocessing step [see Fig. 2(b)] contain either isolated cells or cell clusters. In the last case, the high degree of cell overlap and the inhomogeneities in the nuclei intensity make the detection of the nuclei a difficult task.

Our approach to this problem is based on the gray-scale morphological reconstruction [16] in combination with the detection of regional minima [17] in the image, which are connected components, whose intensity value is the same and less than the intensity value of the external boundary pixels. These minima indicate the positions of the candidate cell nuclei.

Once we have found the regions of cell clusters, we calculate the bounding box containing each cluster and we define the corresponding subimage in the color image. Considering that the nuclei are darker than the surrounding cytoplasm, in each subimage, we search for intensity valleys in the red, green, and blue channels of the color image. These valleys consist of pixels with intensity value lower than a specific threshold, and they are bounded by pixels, whose intensity value is greater than this threshold.

For the formation of homogenous minima valleys, we apply the h-minima transform in the original image [18]. In this way, if the depth of each minimum is greater than or equal to a given threshold h, then the minimum is treated as a marker, otherwise it is eliminated. Thus, shorter peaks are removed, while higher peaks remain, even though they are not as significant as before.

The application of h-minima transform requires the construction of a marker image G, whose peaks determine the location of the objects of interest in the original image. A morphological reconstruction of the original image I from marker G is then performed. For the construction of the marker image G, we subtract a threshold h from every pixel of the complement I of the initial image of dimension D_I

$$G(p) = I(p) - h, \qquad p \in D_I. \tag{3}$$

Following the definition in [16], the gray-scale reconstruction is defined regarding to the elementary geodesic dilation $\delta_I^{(1)}(G)$ of gray-scale image $G \leq I$ "under" I

$$\delta_I^{(1)}(G) = (G \oplus B) \wedge I \tag{4}$$

where $G \oplus B$ is the dilation of G by a flat structuring element B, and $\wedge$ stands for the pointwise minimum. Thus, the gray-scale geodesic dilation of size $n \geq 0$ is obtained by iterating n elementary geodesic dilations

$$\delta_I^{(n)}(G) = \underbrace{\delta_I^{(1)}\left(\delta_I^{(1)}\left(\delta_I^{(1)} \ \ldots \ \left(\delta_I^{(1)}(G)\right)\right)\right)}_{n \text{ times}}. \tag{5}$$

In this equation, the output of an elementary geodesic dilation is used as input in a new elementary geodesic dilation, and this is repeated n times. With the aforementioned definitions, the gray-scale reconstruction $\rho_{\mathrm{I}}(G)$ of image I from marker G is obtained by iterating gray-scale geodesic dilations of G "under" I until stability is reached

$$\rho_I(G) = \lim_{n \to +\infty} \delta_I^{(n)}(G). \tag{6}$$

The algorithm used for the construction of the final image is described in [16]. The final image is the complement of the outcome image and it contains the regional minima, whose depth is less than h, suppressed [see Fig. 3(b)].

For the determination of these regional minima, we perform the nonregional maxima suppression [17] in the complement of the derived image. If we assume that $f(x)$ is the input gray-scale image, F the domain of support for f, and mval the minimum allowed value of f, the output image $g(x)$ is derived as follows:

1. $g \leftarrow f;$
2. $\forall x \in F;$
3. $\quad if\ g(x) \neq \text{mval};$
4. $\qquad if\ \exists y \in \text{Nbr}(x) : g(y) > f(x);$
5. $\qquad\quad g(z) \leftarrow \text{mval}, \forall z \in \Gamma_x \{w : g(w) = f(x)\};$

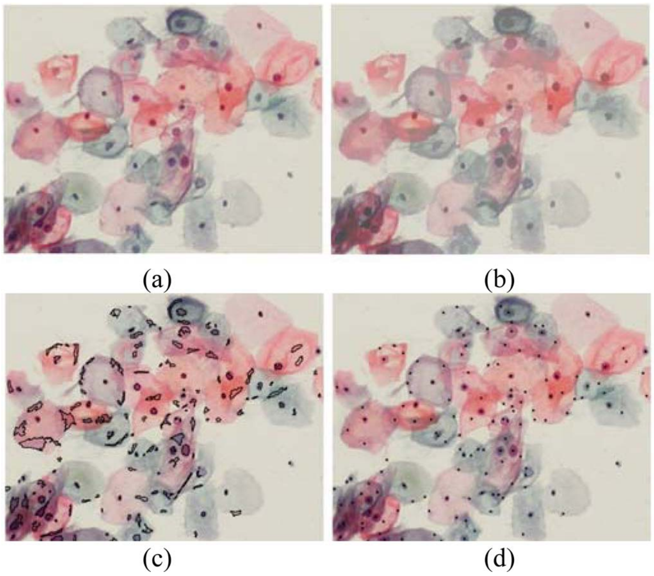


Fig. 3. (a) Initial image of a cell cluster with overlapped cells, (b) resulted image with the suppressed regional minima, (c) areas of regional minima, and (d) centroids of the areas of regional minima.

where Nbr is the neighborhood positions associated with the image position x, and Γ_x is the binary connected opening. This algorithm sets the minimum intensity value to any pixel of the image that does not belong to a regional maximum. If a pixel has a neighbor of higher intensity value, then all pixels connected to this pixel and having the same intensity are set to the minimum allowed value $(\mathrm{mval} = 0)$.

The resulted binary image contains the areas of intensity valleys highlighted. This procedure is independently applied in the three channels of the initial color image obtained after the preprocessing step. The areas of valleys found in the three images are joined using a logical OR operator. Next, the boundaries of these valleys are calculated [see Fig. 3(c)] and the candidate nuclei locations r_c are given by the average of the boundary pixels.

The list of pixels found in this step [see Fig. 3(d)] indicates the locations of the candidate nuclei in the image. However, these centroids do not coincide precisely with the true nuclei centroids, because the boundary of the intensity valleys is rough approximations of the real nuclei boundary. Also, as it can be seen in Fig. 3(d), some undesired points are detected during this step. For the detection of more accurate nuclei centroids and the rejection of unwanted findings, further processing is needed.

C. *Refinement of Candidate Cell Nuclei Centroids*

In this step, *a priori* knowledge about the nucleus appearance is incorporated for the extraction of more accurate nuclei centroids. The nuclei usually have ellipse-like boundaries, from which we can observe that the intensity of the pixels inside these boundaries is lower than those lying outside. As a result, we expect high gradient of the image across the nuclei boundaries.

Nevertheless, the value of the gradient in nucleus/cytoplasm borders varies in different parts of the image This is the reason why edge detectors based on the selection of a threshold in the gradient value are inappropriate for the determination of a more precise nuclei boundary because low thresholds would result in the detection of too many false edges, while high values would result to the loss of some true nuclei boundaries. In this paper, we propose the use of the morphological gradient calculated with an alternative way for the estimation of the nuclei borders.

More specifically, from the initial color image I [see Fig. 4(a)], we construct two different images. The first image A [see Fig. 4(b)] is constructed from the original image I after the application of a gray-scale erosion of the original image, i.e.,

$$A = I \Theta X \tag{7}$$

where X is a flat disk-shaped structuring element with radius 3. The use of a disk-shaped structuring element for the construction of the eroded image pronounce the objects of the image in such a way that dark objects are enlarged radial. The image B [see Fig. 4(c)] is the outcome of the application of a 5×5 averaging filter on the original image. Following this procedure, noise effects and inhomogeneities in nuclei intensity are limited and a smoother image is extracted.

The morphological gradient J of the image I, where the boundaries of the nuclei are accentuated, is defined as follows:

$$J(x, y) = |A(x, y) - B(x, y)|. \tag{8}$$

In this stage, we disregard the color information of the image, as we are interested in the determination of high-intensity differences. For the sharpening of nuclei borders, we apply a contrast enhancement filter in the final image, which saturates 1% of data at low and 1% of data at high intensities of the original image [see Fig. 4(d)]. Finally, in the resulting gradient image, we locally search in each derived centroid for the selection of some points with high-intensity values, which indicate the existence of the nucleus border.

The pixel of the initial candidate nucleus centroid is used as starting point for the construction of a confined search space [see Fig. 4(e)]. The searching area, in which we expect to include the boundary of each nucleus is determined using 8-radial profiles in equal arc length intervals consisted of 8 points each (as this was estimated to be the average size of the nuclei radius by the expert observer). In every radial profile, we choose the pixel with the highest intensity (nonmaximum suppression, [see Fig. 4(f)]. This process is repeated once for each candidate nucleus.

The final step is the redefinition of the nuclei centroids based on the resulted boundary pixels [see Fig. 4(g)]. The outcome of the entire procedure can be observed in Fig. 4(h). This example shows clearly that a more accurate nucleus centroid is detected.

D. *Decision*

The application of the method described previously for the detection of nuclei centroids produces a number of false positive occurrences [see Fig. 3(d)], which must be eliminated. This can be accomplished following a decision process based on two steps: the application of a distance-dependent rule; and the application of classification techniques.

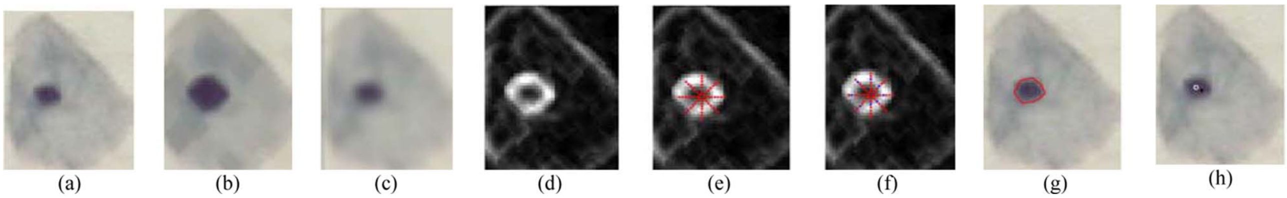


Fig. 4. Illustration of the different steps of the refinement procedure. (a) Initial image, (b) eroded image, (c) filtered image, (d) contrast enhanced image of the difference of images (b) and (c), (e) construction of the search space, (f) determination of pixels in the nucleus circumference by selecting the local maxima of the gradient amplitude, (g) the resulted nucleus contour, and (h) the initial (black cross) and the refined (white circle) centroids of the nucleus.

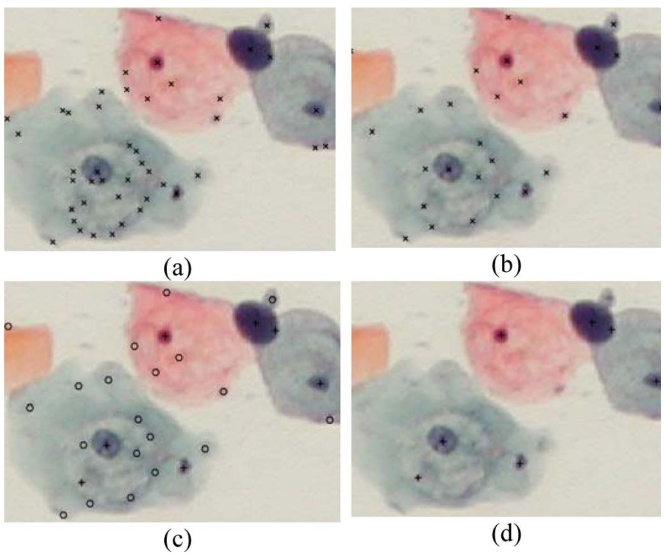


Fig. 5. (a) Initial image with the detected centroids depicted with an "x". (b) Result of the distance-depended rule. (c) Result of the FCM, where the positive (true nuclei) class is depicted with "+" and the negative class (other findings) with "o". (d) Resulted centroids of the positive class.

1) Application of the Distance-Dependent Rule: It is observed that a lot of extracted points are located in small distances between them. Even in the case of one single nucleus, the existence of more than one candidate centroid is possible, and these candidates are generally spread into the nucleus circumference [see Fig. 5(a)]. For this reason, for all the obtained centroids, we apply the following rule:

$$\begin{aligned}
&\text{repeat}\\
&\qquad \forall\, p = (x,y) \in R_c\\
&\qquad \text{if exists}\, q = \{(x_q, y_q) \mid D(p,q) \leq T\}\\
&\qquad\qquad \text{select}\, r = \{p, q \mid \min\{I(p), I(q)\}\}\\
&\qquad\qquad \text{update}\, R_c\\
&\text{until no change in}\, R_c
\end{aligned}$$

where R_c is the set of all centroids, D is the Euclidean distance between two points, T is the threshold on the minimum distance, and $I(p)$ is the intensity of the image at point p. The threshold for the minimum distance that we use is derived from the prior knowledge we have about the true diameter of a nucleus. By applying this rule, we have a significant reduction of the total number of the resulted centroids, while at the same time, we have no loss of the true nuclei [see Fig. 5(b)].

2) Application of Classification/Clustering Techniques: In the final set of the candidate nuclei centroids, we proceed with the application of classification algorithms for the separation of the points of true nuclei and the points that belong to other regional minima. We have tested our method using an unsupervised and a supervised classification algorithm, namely the fuzzy C-means (FCM) [19] and the support vector machine (SVM) [20], respectively. Given the fact that the FCM algorithm does not require any training, it is independently applied in each image. Representative results of the FCM clustering algorithm in the real image are shown in Fig. 5(c)–(d).

For the application of the SVM classification algorithm, a training data set is constructed by random selection of 34 images from the entire data set. The remaining four images are used as test set. This experiment was repeated 20 times, each time using a different (randomly selected) training set. After training, the performance of the SVM classifier is calculated using the unknown images of the test set. It must be noted that in our experiments, we have used the linear and the radial basis function (RBF) kernels.

3) Feature Vectors: For the definition of the set of nuclei patterns, we have used the intensity information of the neighborhood of the centroids. We have tested the performance of our method using four pattern sets of different sizes for the neighborhood, that is D1 with $3 \times 3 \times 3$ pattern size, D2 with $5 \times 5 \times 3$ pattern size, D3 with $7 \times 7 \times 3$ pattern size, and D4 with $9 \times 9 \times 3$ pattern size (the third dimension corresponds to the color). Each pattern was centered at each centroid in the initial color image. We have constructed two data sets of patterns using as the center of the neighborhood the initial and the refined centroids, respectively.

III. Results

A. Study Group

Our data set is composed by 38 conventional Pap-stained cervical cell images from 15 different Pap smear slides, acquired through a CCD camera (Olympus DP71) adapted to an optical microscope (Olympus BX51) using a $10\times$ magnification lens. The size of the images is 1536×2048 and they were stored in JPEG format. The total number of cell nuclei in the images is 5617. In order to obtain the ground truth, the nuclei locations were manually identified by two expert cytopathologists. The

TABLE II
EXECUTION TIME OF THE PROPOSED METHOD FOR IMAGES OF SIZE 1536 × 2048

Step of the proposed method	Time in sec. (mean±std)
Preprocessing	3.53 ± 0.22
Detection of candidate cell nuclei centroids	72.45 ± 39.55
Distance dependent rule	7.06 ± 13.00

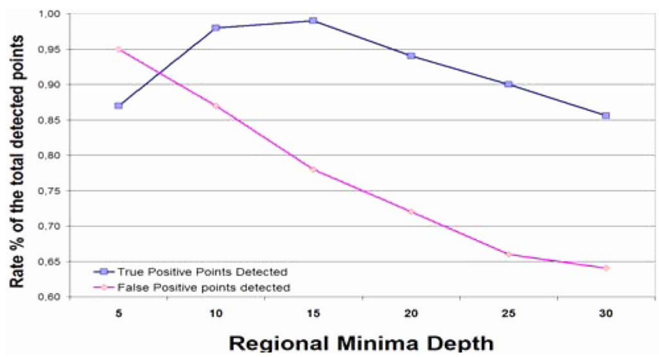


Fig. 6. Rate of the true positives (true nuclei centroids detected) and false positives for different thresholds in regional minima depth.

types of the existed cell nuclei are all the aforementioned in Section I, and there also exist some abnormal nuclei.

B. Numerical Evaluation

For the evaluation of the performance of the method, we have to examine the performance of the different steps of the method. Furthermore, as a measure of the computational efficiency of the segmentation method, we present in Table II the processing times of the individual steps of the method developed in MATLAB using a Pentium 2.0 GHz with 3 GB RAM.

The preprocessing is a fast procedure that results in the determination of the parts of the image containing isolated cells or cell clusters. It misses nine cell nuclei in all images and it produces a reduction of true positives cell nuclei of 0.16% of the total initial number of nuclei. The loss of this step is mainly due to the existence of some faintly stained cell cytoplasms, which are not distinguishable from the background. Thus, the nucleus is removed, as it is considered to be an isolated object.

The detection step of the cell nuclei centroids successfully identifies most of the nuclei in the image. In this step, 42 true nuclei are missed and the true nuclei detection rate is 99.25%. For the choice of the threshold of the depth of the intensity valleys, we have performed several tests, and as it is depicted in Fig. 6, with the threshold value of 15, we obtain the maximum number of true nuclei centroids detected.

The distance-dependent rule on the refined nuclei centroids yields in the reduction of false positive findings at the rate of 14.13%, while we have no loss of true nuclei centroids. This rate could be higher if we select a distance threshold higher than 8 pixels. However, with a selection of a higher value for this threshold, true nuclei centroids are missed, as it can be observed in Fig. 7. It must be noted that if we omit this step, there will be some centroids, which belong to the same nucleus and they will introduce interference in the clustering step. For instance,

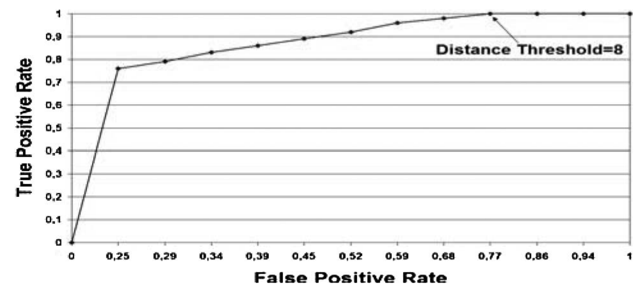


Fig. 7. ROC curve used for the threshold selection in distance-depended rule.

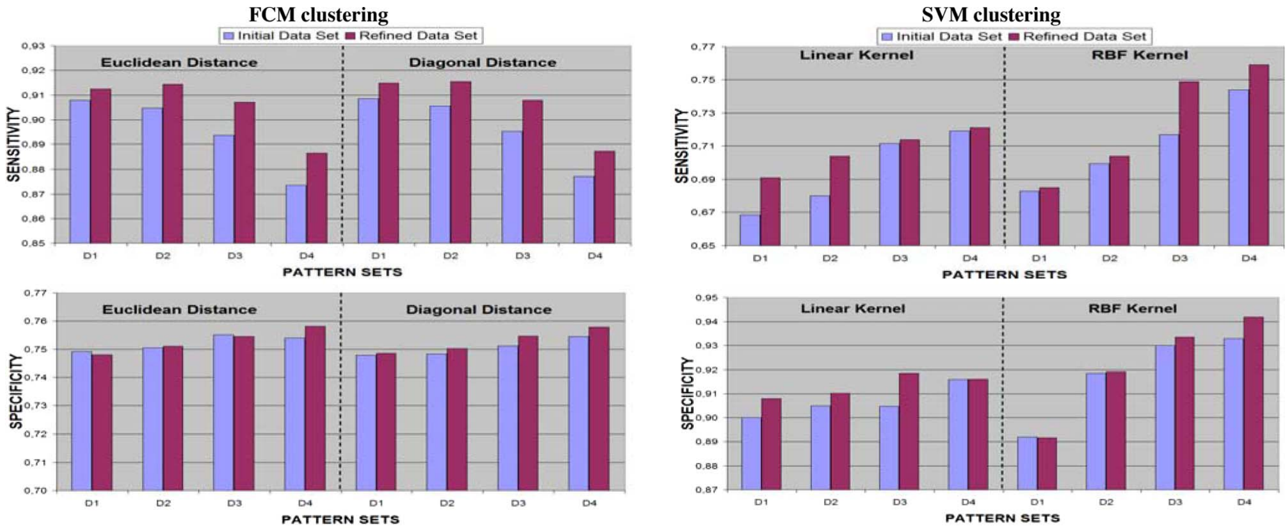
if they are classified in the same class (e.g., the nuclei class), we will not be able to compute the number of true detected nuclei, since one single detected nucleus will be counted twice. Furthermore, if they are assigned to different classes, then one centroid will be counted as true positive and the other one as false negative. This would be wrong since both belong to the same nucleus.

For the application of the classification algorithms, we have used two data sets, as it is already described. In FCM algorithm, we have used the Euclidean and the diagonal norm as the distance-dependent metric. The Euclidean norm between vectors u and v of dimension N, is defined by $D^{\mathrm{Euc}}(u,v) = \sqrt{(u-v)^T(u-v)}$. Respectively, the diagonal norm is defined by $D^{\mathrm{Diag}} = \sqrt{(u-v)^T A_D (u-v)}$,where A_D is a diagonal matrix containing the standard deviations of the vectors. Furthermore, the SVM classifier leads to the selection of some tens of support vectors, depending on the type of kernel, the data set, and the dimensions of the patterns that we use.

For the comparison of the results, we have calculated two widely used statistical measures, the *sensitivity* and the *specificity* (our images are annotated, and the true positive and false positive findings are automatically determined). As it is depicted in Fig. 8(a) and (b), the FCM has higher *sensitivity* rate than the SVM, which means that fewer true nuclei are missed. However, the *specificity* of FCM is low relatively to SVM, which means that FCM includes a lot of false positive centroids in the final set of the points characterized as nuclei centroids by the algorithm. On the contrary, the *sensitivity* of the SVM classification is relatively low, namely, it misses more true nuclei centroids. Nevertheless, it presents high *specificity* rate, which means that in the final set of points characterized as nuclei, the false positives are limited. An important fact that must be noted is that in both FCM and SVM, the use of the refined centroid data set leads to a better classification performance. This is explained by the fact that the refined centroids are closer to the true nuclei centroids and the produced patterns contain more representative features of the nuclei, which results in the improvement of the discrimination ability of the classification techniques.

IV. DISCUSSION

The proposed method is fully automated and its application was performed without any observer interference. The parameters of the several steps of the method (see Table III) were



(a) (b)

Fig. 8. Results of the application of the (a) FCM clustering and the (b) SVM clustering with respect to sensitivity and specificity.

TABLE III
VALUES OF THE PARAMETERS OF THE PROPOSED METHOD

Step of the method	Parameter	Value
Preprocessing	Area threshold	500
Detection of candidate cell nuclei centroids	Intensity depth threshold (h)	15
	Minimum allowed Image value ($mval$)	0
Refinement of candidate nuclei centroids	Number of radial profiles	8
	Length of radial profiles	8
Distance Dependent Rule	Minimum Distance threshold (T)	8
FCM	Weighting component (m)	2
SVM	Linear Kernel: C	0.001
	RBF Kernel: C	1
	RBF Kernel: γ	1

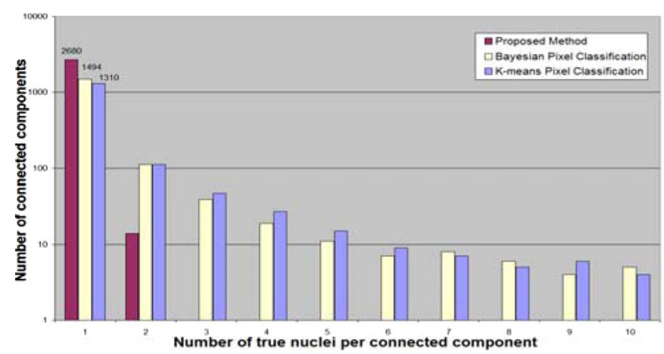


Fig. 9. Comparative results of our method and the pixel-classification schemes proposed in [5] in terms of correct nuclei localization.

computed after several experiments in 19 randomly selected images from the entire data set containing 3616 images, and afterward the method was applied in all 38 images of our data set.

In Table II, the processing times of the individual steps of the method are provided. As we can see, the execution time significantly varies. In the regional minima step, the number of the real cell nuclei in each image affects the execution time, and since our images contain 26 –522 nuclei, the method exhibits high variation in the execution time of this step. Furthermore, the proportion of the image that is identified as background in the preprocessing step is another factor that influences the execution time. In an image with artifacts, severe noise and variation in cell staining, the preprocessing step results in the selection of some regions of the image that do not correspond to the true location of the cell clusters. Even though in those areas no cells are present, the regional minima detection step is also performed in those areas, and this demands additional execution time. This also results in the detection of false positive findings, which affects the execution time of the distance-dependent-rule step, because more candidate points are processed. Finally, the variation of the execution time of these steps is affected by the presence of outlying images that exhibit high difference from the mean execution time, and although are a few, their influence in the total variation is significant.

Considering the classification performances, the selection of one of the classification techniques (FCM or SVM) depends on the purpose of the detection of nuclei in a specific Pap smear image. For instance, if the purpose is to find abnormal or malignant cell nuclei, the FCM is preferable, as it produces lower loss of true nuclei and the probability of a missed abnormal nucleus is reduced. On the other hand, if the purpose is to detect cells nuclei in order to calculate, for example, morphological characteristics, a pure set of true nuclei would be desirable and the SVM classification technique is suitable, as it reduces the false positive occurrences in the final set. However, since the performance of SVM depends on the selected values of the parameters, its use becomes more demanding, especially when a limited number of images exist. On the other hand, the application of the FCM algorithm can be performed directly in one single image. As a result and in combination with the high performance it presents, the FCM algorithm is preferable for the classification step of our method.

It must be noted that the artifacts and background nonuniformities make the detection of false positives (and the existence of two classes) highly probable. However, in the extreme and rare case of having only true positive findings in the classification step there are two possibilities. If the nuclei are homogeneous,

TABLE IV
COMPARISON OF THE PROPOSED METHOD AND OTHER METHODS APPEARED IN THE LITERATURE

Method	Smear slides	Images	Image size	Cells	Performance Criteria	Quantitative Results
Bamford *et al.* [6]	Unknown	Unknown	128×128	20130 (1 cell/ image)	Visual inspection.	99.64% correctly segmented cells.
Wu *et al.* [7]	1	1	80×100	1	Comparison with K-means and Bayes classifier in a synthetic image.	Misclassification rate lower than 5%.
Garrido *et al.* [8]	Unknown	3	Unknown	Unknown	Visual inspection.	Lack of quantitative results.
Lezoray *et al.*.[5]	Unknown	10	Unknown	209 manually segmented regions	Vinet measure. Number of segmented regions.	Vinet measure:2.24 (RGB) - 3.41 (HSL). Mean difference from the ground truth: 2.87%(RGB) - 0.47% (HSL).
Lassouaoui *et al.* [9]	Unknown	2	256×256	Unknown	Visual inspection.	Lack of quantitative results.
Bak. *et al.* [3]	Unknown	2	Unknown	Unknown	Visual inspection.	Lack of quantitative results.
Mat Isa. *et al.* [10]	Unknown	3	Unknown	Unknown	Visual inspection.	Lack of quantitative results.
Yang-Mao *et al.* [12]	Unknown	Unknown	64×64	124 (1 cell/image)	Misclassification error,edge mismatch, relative foreground area error, modified Hausdorff distance, region nonuniformity, relative distance error.	Average segmentation error for the nuclei of 0.1145.
Lin *et al.* [11]	Unknown	10	Unknown	10 (1 cell/ image)	Misclassification error, relative foreground area error, modified Hausdorff distance.	Average segmentation error for the nuclei of 0.1323.
This Work	15	38	1536×2048	5617	Sensitivity (Se), Specificity (Sp)	Indicative mean values: FCM: 90.57% (Se), 75.28%(Sp) SVM 69.86% (Se), 92.02% (Sp)

they will be assigned to the same class. If the nuclei exhibit dissimilarities, then they would be split into two classes. Moreover, in practice, the true positive class should be indicated by the user at the end of the procedure.

We have also compared our method with the detection methods proposed by the state of the art technique of Lezoray and Cardot [5], which is based on the k-means clustering algorithm and a Bayesian pixel-classification scheme. Following the principles in [5], these schemes classify each pixel of the images (with the background removed) as "nuclei" or "cytoplasm" pixel.

All the parameters of the mixture of Gaussian distributions were calculated on a training set of color vectors from randomly selected images of our data set (50% of the images). Then, the Bayesian classifier was applied in the remaining images. This experiment was repeated five times, each time with a different (randomly selected) training set.

The outcome of both pixel-classification schemes are connected components of probable nuclei locations and they are compared with the outcome of the detection of regional minima step in terms of how many true nuclei centroids were recognized. The expected results would be the detection of one nucleus per connected component. Thus, the desirable performance of each method is a high number of connected components containing only one nucleus. In Fig. 9, we can observe the average number of the detected connected components, over the test sets of images, which were recognized by the compared methods. As we can see, our method is superior to the pixel-classification schemes, since it produces more single connected components, which contain only one nucleus. Let us also notice that the vertical axis in Fig. 9 has a logarithmic scale making the differences in performance more pronounced.

Beyond the comparison of our method with pixel-classification schemes, Table IV shows a comparison of our method and other methods appeared in the literature. In general, it is difficult to compare the methods directly since many of them do not include quantitative results and the performance criteria extensively vary. Furthermore, some data parameters are not clearly defined, which are important for the evaluation of the general behavior of each method.

From Table IV, we can assert that our method is superior for several reasons. First, the data set that was used includes images captured from 15 different Pap smear slides, which evince that the data set contains a big variety of different cells, and the obtained results describe more precisely the general behavior of the method and the expected performance in a new image. Also, the proposed method can be applied in images captured directly from an optical microscope and is able to successfully recognize the cells nuclei, even in cases, where cell overlapping is present. Moreover, the average number of cells nuclei in these images is 148, and they are clearly more complicated than the images containing only isolated cells, such as in [4], [6], [9], [11], and [12].

In terms of the general image-processing approach, the method exploits the color information of the image, in contrary to the techniques in [4], and [6]–[12]. This is advantageous, since the staining process of the smear has different effects in the three-color components of the image and some nuclei are more distinguishable in a single-color channel. The use of three different thresholds (one for each color channel) in the Otsu's method in the preprocessing step is more effective than the use of one single threshold in the gray-scale image. Furthermore, the detection of the intensity valleys in the three channels of a color image and the merge of the detected regions in a final image results in the determination of more true nuclei locations, rather than the detection of the intensity valleys in the gray-scale image. As we can see in Fig. 10, both the preprocessing and the regional minima step fail to recognize the same number of the true nuclei in the gray-scale image. The individual processing of each color component and the combination of the results leads in no loss of information. However, an issue that must be solved in the future is the recognition of clustered and abnormal nuclei.

V. CONCLUSION

The task of identifying the cell nuclei in conventional Pap smear images is a challenging issue. We have developed a

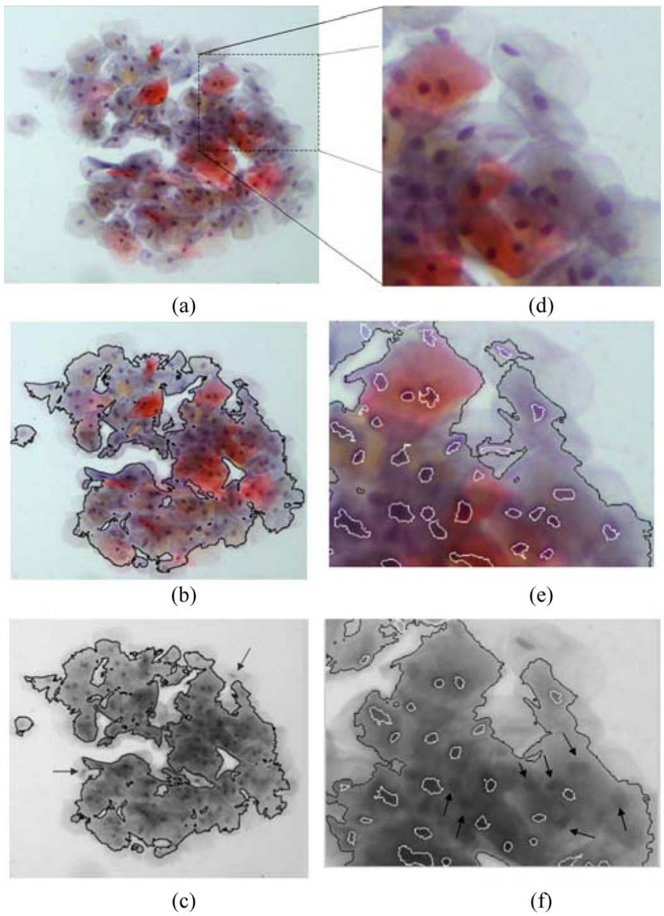


Fig. 10. (a) Initial image. (b) Result of the preprocessing step (denoted with the black line) in the color image. (c) Corresponding result in the gray-scale image. (d) Part of the initial image (e) Result of the detection of regional minima step (denoted with white lines) in the color image. (f) Corresponding result in the gray-scale image. The missed nuclei in the gray-scale images are marked with the arrows in both cases.

robust and accurate method for the automated identification of the cells nuclei, which can be used as the basis for further processing of cell images. As our image data set derives from different Pap smear slides, the method is expected to present high performance, when it is applied in a new Pap smear image. The major advantage of the proposed method is that it is fully automated and it is suitable for images with high degree of cell overlapping, as it can successfully detect not only the nuclei of isolated cells, but also the nuclei in cell clusters.

References

[1] G. N. Papanicolaou, “A new procedure for staining vaginal smears,” *Science*, vol. 95, no. 2469, pp. 438–439, 1942.

[2] H. S. Wu, J. Gil, and J. Barba, “Optimal segmentation of cell images,” *IEE Proc. Vis., Image Signal Process.*, vol. 145, no. 1, pp. 50–56, Feb. 1998.

[3] E. Bak, K. Najarian, and J. P. Brockway, “Efficient segmentation framework of cell images in noise environments,” in *Proc. 26th Int. Conf. IEEE Eng. Med. Biol.*, Sep., 2004, vol. 1, pp. 1802–1805.

[4] P. Bamford and B. Lovell, “A water immersion algorithm for cytological image segmentation,” in *Proc. APRS Image Segmentation Workshop*, Sydney, Australia, 1996, pp. 75–79.

[5] O. Lezoray and H. Cardot, “Cooperation of color pixel classification schemes and color watershed: A study for microscopic images,” *IEEE Trans. Image Process.*, vol. 11, no. 7, pp. 783–789, Jul. 2002.

[6] P. Bamford and B. Lovell, “Unsupervised cell nucleus segmentation with active contours,” *Signal Process.*, vol. 71, no. 2, pp. 203–213, 1998.

[7] H. S. Wu, J. Barba, and J. Gil, “A parametric fitting algorithm for segmentation of cell images,” *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 400–407, Mar. 1998.

[8] A. Garrido and N. P. de la Blanca, “Applying deformable templates for cell image segmentation,” *Pattern Recognit.*, vol. 33, no. 5, pp. 821–832, 2000.

[9] N. Lassouaoui and L. Hamami, “Genetic algorithms and multifractal segmentation of cervical cell images,” in *Proc. 7th Int. Symp. Signal Process. Appl.*, 2003, vol. 2, pp. 1–4.

[10] N. A. Mat Isa, “Automated edge detection technique for Pap smear images using moving K-means clustering and modified seed based region growing algorithm,” *Int. J. Comput. Internet Manag.*, vol. 13, no. 3, pp. 45–59, 2005.

[11] C. H. Lin, Y. K. Chan, and C. C. Chen, “Detection and segmentation of cervical cell cytoplast and nucleus,” *Int. J. Imaging Syst. Technol.*, vol. 19, no. 3, pp. 260–270, 2009.

[12] S. F. Yang-Mao, Y. K. Chan, and Y. P. Chu, “Edge enhancement nucleus and cytoplast contour detector of cervical smear images,” *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 38, no. 2, pp. 353–366, Apr. 2008.

[13] M. E. Plissiti, E. E. Tripoliti, A. Charchanti, O. Krikoni, and D. Fotiadis, “Automated detection of cell nuclei in Pap stained smear images using fuzzy clustering,” in *Proc. 4th Eur. Congr. Med. Biomed. Eng.*, 2008, vol. 22, pp. 637–641.

[14] K. Zuiderveld, “Contrast limited adaptive histogram equalization,” in *Graphics Gems IV*. San Diego, CA: Academic, 1994, pp. 474–485.

[15] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

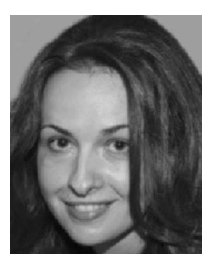
[16] L. Vincent, “Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms,” *IEEE Trans. Image Process.*, vol. 2, no. 2, pp. 176–201, Apr. 1993.

[17] E. J. Breen and R. Jones, “Attribute openings, thinning, and granulometries,” *Comput. Vis. Image Understanding*, vol. 64, no. 3, pp. 377–389, 1996.

[18] P. Soille, *Morphological Image Analysis: Principles and Applications*. New York: Springer-Verlag, 1999.

[19] J. C. Bezdek and S. K. Pal, *Fuzzy Models for Pattern Recognition*. New York: IEEE Press, 1992.

[20] N. Christianini and J. S. Taylor, *Support Vector Machines and Other Kernel-Based Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.



Marina E. Plissiti received the B.Sc. and M.Sc. degrees from the Department of Computer Science, University of Ioannina, Greece, in 1998 and 2001, respectively. She is currently working toward the Ph.D. degree in the same Department.

Since 2001 she is a Secondary School Teacher. Her research interests include medical image processing and artificial intelligence in biomedical applications.

Christophoros Nikou received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1994 and the DEA and Ph.D. degrees in image processing and computer vision from Louis Pasteur University, Strasbourg, France, in 1995 and 1999, respectively.

During 2001, he was a Senior Researcher with the Department of Informatics, Aristotle University of Thessaloniki. From 2002 to 2004, he was with Compucon S.A., Thessaloniki. Since 2004, he is with the Department of Computer Science, University of Ioannina, Greece where he was a Lecturer (2004-2009) and since 2009, he has been an Assistant Professor. His research interests mainly include image processing and computer vision and their application to medical imaging.

Antonia Charchanti, photograph and biography not available at the time of publication.