Watershed-based segmentation of cell nuclei boundaries in Pap smear images

Marina E. Plissiti, Christophoros Nikou, and Antonia Charchanti

***Abstract*—In this work we present a fully automated method for the accurate detection of cell nuclei boundaries in conventional Pap smear images, based on the watershed transform. For the extraction of nuclei and cytoplasm markers, which are used as starting points for the flooding process, a morphological reconstruction step is initially performed in each image. The watershed transform is then applied in the color morphological gradient image, which shows the boundaries of the more pronounced nuclei. For the elimination of false positive findings, salient features of shape and intensity of the detected regions were calculated and a clustering step is performed. The method was evaluated with a data set of 19 images containing 3616 recognized cells nuclei. The performance of the method was evaluated in terms of the correct detection of the positions of the nuclei. Comparisons with the segmentation results of the gradient vector flow (GVF) deformable model showed that the segmentation of the watershed transform captures more accurately the boundaries of nuclei, leading to a better performance of the clustering algorithm.**

I. Introduction

ONE of the most interesting fields in cytological image analysis is the automated detection of cells nuclei in Pap smear images. These images depict exfoliated cells from the epithelium which are stained using a staining technique introduced by G. Papanicolaou [1]. Nowadays, the Pap smear is extensively used in gynaecology as a screening test for the detection of premalignant and malignant processes in the cervix. The interpretation of these images relies on the visual recognition of changes in the nuclei of cells affected by disease. However, this process is a tedious, time-consuming and in many cases error-prone procedure due to the high degree of complexity that these images exhibit.

Several approaches have been proposed for the segmentation of Pap smear images and they concern techniques such as active contours [2], template fitting [3] edge detectors [4], pixel classification schemes [5], genetic algorithms [6] and region growing with moving K-means [7].

In our work, we propose a fully automated method for the segmentation of cell nuclei based on the watershed transform, which can be applied directly in Pap smear images containing both isolated and clustered cells (Fig. 1(a)). A segmentation process and a clustering step are combined resulting in the determination of the nuclei boundaries and the elimination of false positive findings, respectively.

More specifically, at first in the segmentation process, the centroids of the regional minima of the image are detected through a morphological reconstruction step [8], and afterwards the boundaries of the nuclei are defined with the watershed transform. It must be noted that the watershed transform is performed in the morphological color gradient image [9], in order to exploit the color information of the image in the estimation of the nuclei borders. The method overcomes the problem of the determination of nuclei and cytoplasm markers, as they are defined automatically. From the detected markers, a flooding process begins, in order to avoid the oversegmentation, which usually occurs when the watershed transform is applied to an image without markers.

The clustering step is then performed using Fuzzy C-means algorithm, with features extracted from the regions enclosed from the final boundaries, obtained from the watershed transform. These features concern the intensity of the enclosed area and the shape of the final boundary. This step is essential for the elimination of the detected regions that do not correspond to the true nuclei locations.

The accuracy of the segmentation was compared with the corresponding segmentation of the Gradient Vector Flow (GVF) [10] deformable models, in terms of correct detection of true nuclei regions. As it was verified by the results, the performance of the classification method is higher when the features of the area enclosed by the boundaries obtained by the watershed transform are used. This implies that the watershed transform results in more accurate nuclei boundaries than GVF segmentation. The proposed method is fully automated and it can be applied in any microscopic cervical cell image.

II. Materials and Method

A. Study Group

We have collected 19 images of conventional PAP stained cervical cell slides, which were acquired through a CCD camera adapted to an optical microscope. We have used a 10× magnification lens and the acquired images were stored in JPEG format. The total number of cell nuclei in the images was identified by an expert observer and they are in total 3616.

Manuscript received July 5, 2010.

M. E. Plissiti is with the Computer Science Department, University of Ioannina, GR 45110, Ioannina, Greece (e-mail: marina@cs.uoi.gr).

C. Nikou is with the Computer Science Department, University of Ioannina, GR 45110, Ioannina, Greece (phone: +30-26510-08802; fax: +30-2651-0-08880; e-mail: cnikou@cs.uoi.gr).

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

978-1-4244-6561-3/10/$26.00 ©2010 IEEE

B. Method

The prerequisite for the avoidance of oversegmentation that the watershed transform produces is the detection of nuclei and cytoplasm markers, from which the flooding process will begin.

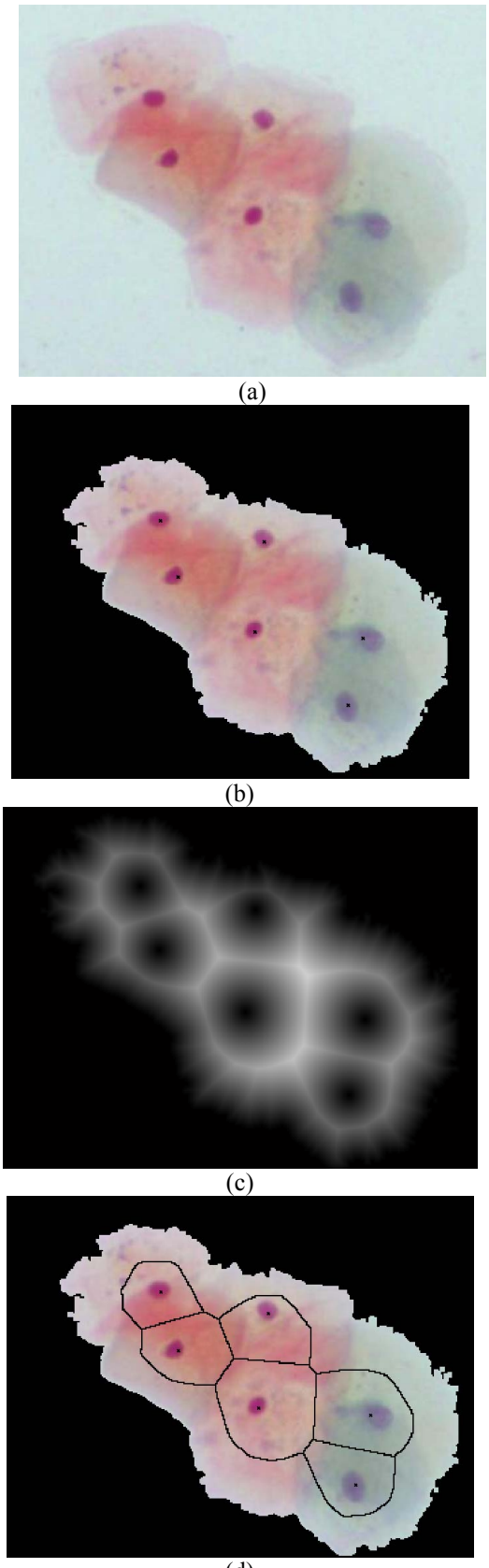
1) Detection of candidate nuclei markers

The aim of this step is to define the nuclei markers in the images, which will be used as starting points in the flooding process of the watershed transform, for the determination of the boundaries of the nuclei. In a first step, each Pap smear image is preprocessed in order to extract the background and to define the regions of interest (Fig. 1(b)). In the preprocessing step we perform sequentially the contrast-limited adaptive histogram equalization and global thresholding with the standard method proposed by Otsu [11] to the red, green and blue component of the image, which result in three binary images. The final outcome of this step is a binary mask, which is the result of a logical OR operation of these three binary images. In this mask, all the connected components with an area smaller than the area of a single isolated cell are removed. This is necessary for the exclusion of image artifacts that may interfere in the next steps.

In the regions of the image found in the preprocessing step a morphological process is performed. Considering that the nuclei are darker than the surrounding cytoplasm, we search for intensity valleys in the image. For this reason, the *h*-minima transform [12] in the red, green and blue components of the original image is applied, for the formation of homogenous minima valleys. A morphological reconstruction step is then performed. In this step, we use the resulted image as a mask and the initial image as a marker. In the final image, the regional minima whose depth is less than *h* are considered to correspond to the nuclei areas and the location of each candidate nucleus is determined with the centroid. For the elimination of two or more detected markers in an area of a radius that it is smaller than the mean radius of a normal nucleus, a distance dependent rule is performed, as it is described in [8].

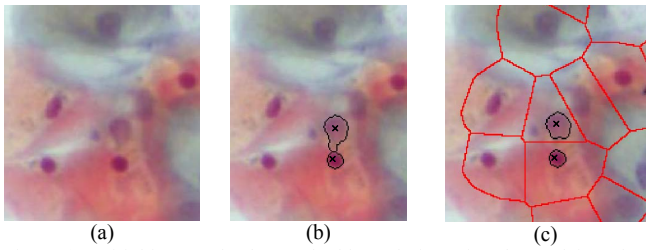
2) Detection of cytoplasm markers

After the detection of the nuclei markers, some cytoplasm markers are necessary for the definition of the regions in the image, which correspond to the catchment basins that each nucleus belongs to. This step is essential for the extraction of accurate nuclei boundaries, since the use of the nuclei markers alone may result in noisy boundaries (Fig 2(b)). For this reason, we perform the distance transform in the binary mask obtained in the preprocessing step, with the nuclei markers superimposed (Fig. 1(c)). In the resulting image, the watershed transform is performed and the watershed lines separate the image into distinct areas. The flooding process, which will start from the nuclei markers, will be restricted in these areas, preventing the detected nucleus boundary to merge with adjacent nuclei borders (Fig. 2(c)).



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

Fig. 1: (a) Initial Pap smear image, (b) the binary mask with the background excluded and the nuclei markers superimposed, (c) the result of the distance transform, (d) the nuclei markers (depicted with "x") and the cytoplasm markers (depicted with the black line).



(a) (b) (c)

Fig. 2: (a) Initial image, (b) the resulted boundaries using the nuclei markers only and (c) the segmentation of the watershed transform using both nuclei and cytoplasm markers (depicted with red lines).

3) Definition of the nuclei boundaries

A powerful technique for the delineation of objects of interest is the watershed transform. The concept of watersheds is based on the consideration that the image in three dimensional space, is described with two spatial coordinates versus intensity, which is assumed to be the elevation information. Given this topographic representation of the image, the pixels are divided into separate catchment basins, which are determined by the regional minima of the image or a predefined set of markers. The aim of this technique is to determine the pixels of the watershed lines, which separate neighboring catchment basins and consequently they separate different characteristic parts of the image.

For the application of the watershed transform, an image containing pronounced nuclei boundaries is required. Usually most of the cells nuclei have ellipse-like shape and they consist of pixels whose intensity is lower than those pixels lying outside the nuclei. As a consequence, high gradient of the image across the nuclei boundaries is expected. For this reason, and in order to combine the different influence of each color component of the image in the nucleus boundary, we construct the color morphological gradient image [9]. In general, the color morphological gradient is an extension of the morphological gradient of a grayscale image in color space. More specifically, the morphological gradient is defined in terms of grayscale dilation $\delta_g(f)$ and grayscale erosion $\varepsilon_g(f)$ as:

$$\nabla(f) = \delta_g(f) - \varepsilon_g(f) \,, \tag{1}$$

Another way for the expression of the morphological gradient is the maximum absolute intensity difference between two pixels in the area of the structuring element G, that is:

$$\begin{aligned} \nabla(f) &= \max_{x\in G}\{f(x)\} - \min_{x\in G}\{f(x)\} \\ &= \max\left(\left|f(x) - f(y)\right|\right) \ \forall x, y \in G \end{aligned} \tag{2}$$

In color images, the pixels are denoted as three dimensional vectors with elements the intensities in the three color components and the color morphological gradient (CMG) can be expressed as:

$$\text{CMG} = \max_{i,j\in G}\left\{\left\|x_i - x_j\right\|_p\right\} \tag{3}$$

where x_i, x_j are pixels in the structuring element G. In our experiments the second norm $(p=2)$ was calculated and a 3×3 flat structuring element was used. In the obtained image, the watershed transform with the detected nuclei and cytoplasm markers is performed, resulting in the definition of the nuclei boundaries. However, some false positive findings are detected (Fig 3(a)) and a clustering step is necessary for the final definition of the nuclei set.

TABLE I
INTENSITY AND SHAPE FEATURES

Feature	Formula
Average image intensity in Red	$AVG_R = \sum_{i=1}^{N} I_R(x_i, y_i)/N$
Average image intensity in Green	$AVG_G = \sum_{i=1}^{N} I_G(x_i, y_i)/N$
Average image intensity in Blue	$AVG_B = \sum_{i=1}^{N} I_B(x_i, y_i)/N$
Diameter of a circle with the same area as the region	$DM = 2\sqrt{Area/\pi}$
Proportion of the convex hull pixels that are also in the region	$CV = Area/ConvexArea$
Major axis length[(1)]	$L = \sqrt{\frac{2\left(u_{20}+u_{02}+\sqrt{4u_{11}^2+(u_{20}-u_{02})^2}\right)}{u_{11}}}$
Minor axis length[(1)]	$K = \sqrt{\frac{2\left(u_{20}+u_{02}-\Delta=\sqrt{4u_{11}^2+(u_{20}-u_{02})^2}\right)}{u_{11}}}$
Eccentricity	$E = 2\frac{\sqrt{\left(\frac{L}{2}\right)^2-\left(\frac{K}{2}\right)^2}}{L}$

[(1)] Major and minor axis length are calculated for an ellipse that has the same second moments (u_{pq}) as the region.

4) Clustering of the candidate nuclei areas

In general, the cells nuclei present significant variations in Pap smear images, which mainly include differences in their color intensity and their size. The construction of a compact set of characteristic features of the nuclei would contribute in the correct detection of the nuclei in these images.

For this reason, eight features for each candidate nucleus area are calculated, which concern the intensity and the shape attributes of the region enclosed by the detected boundaries (Table I). Then the Fuzzy C-means clustering algorithm is performed, for the separation of the segmented regions of the image, into two classes (the true nuclei class and the other findings class). As it was verified by the results, the features calculated from the watershed segmentation are more representative for the true nuclei class, since the clustering algorithm has higher performance in comparison with the GVF segmentation.

III. RESULTS AND DISCUSSION

The method was applied automatically in 19 images which were captured from different Pap smear slides. In the detection of the candidate nuclei markers step, the method successfully recognizes 99.39% of the total nuclei in all images. Some of the true nuclei are missed in this step, and this is mainly due to the fact that some of the cells are faintly stained and they are undistinguished from the background. In that case, the nuclei of these cells are considered as isolated objects in the image background and they are rejected as image artifacts. Furthermore, the uneven layering of some cells results in areas of low intensity and the intensity of the nucleus does not well differentiate from the cytoplasm

TABLE II
CLASSIFICATION RESULTS

	Sensitivity	Specificity	HM
Watersheds	88.92%	79.46%	83.92%
GVF	91.77%	74.23%	82.07%

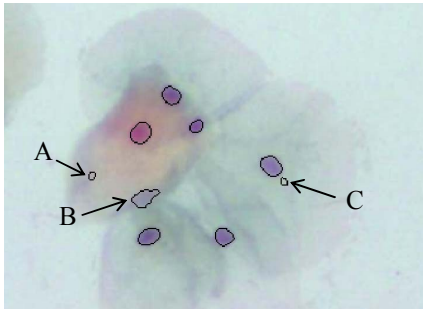
intensity. As a consequence, no regional minimum is detected in the nucleus position.

The boundaries extracted from the watershed transform are used for the calculation of intensity and shape features of each detected region, in order to exclude false positive occurrences through a clustering step. Furthermore, the GVF deformable models were also performed as in [8] and the same features were calculated from the resulted boundaries. These features are used as input in the Fuzzy C-means clustering algorithm.

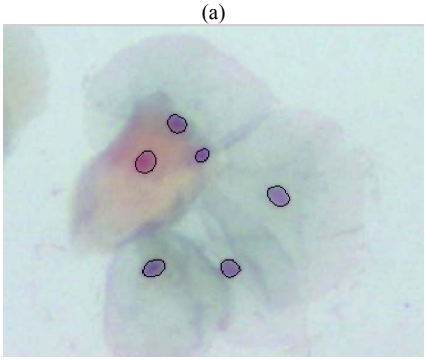
Two widely used statistical measures for the performance of the classification and their harmonic mean (HM) are calculated: the sensitivity, which measures the proportion of actual nuclei which are correctly identified as such by the classification algorithm, and the specificity, which measures the proportion of the detected regions that are not nuclei and are correctly characterized as such. The results of the Fuzzy C-means clustering algorithm are presented in Table II. As we can see, although the sensitivity of the GVF segmentation is higher than the watersheds, the overall performance which is expressed by the harmonic mean (HM) is higher for the watershed segmentation. Given the fact that the accurate determination of the nuclei boundaries leads to the calculation of reliable features, we can conclude that the watershed segmentation entails in more accurate nuclei boundaries than the GVF segmentation. This is confirmed by the high performance of the clustering algorithm.

IV. CONCLUSION

The correct identification of the cell nuclei areas in Pap smear images is a prerequisite for the derivation of diagnostic decisions. We have developed a method for the segmentation of nuclei boundaries in Pap smear images which is fully automated and it can be applied directly in any conventional Pap stained cervical smear image. The method overcomes the problem of the definition of the nuclei and cytoplasm markers, in order to avoid the oversegmentation that the watershed transform would produce. As it was verified by the results, the outcome of the watershed transform is accurate nuclei boundaries, since the performance of the clustering algorithm is high, using a feature set calculated by the enclosed regions of the boundaries. The future work includes efforts to improve the performance of the clustering algorithm with the selection of different nuclei features. Furthermore, the method is being tested in a larger data set in order to evaluate the robustness of the method.



(a)



(b)

Fig. 3: (a) The result of the watershed transform in clustered cells. Notice that the regions A, B and C are false positive findings, due to the detection of regional minima of the image that do not correspond to the true nuclei positions. (b) The final segmentation results after the clustering step.

REFERENCES

[1] G. N. Papanicolaou, "A new procedure for staining vaginal smears", Science, 95 (2469), 1942, pp. 438-439.G.

[2] P. Bamford, B. Lovell, "Unsupervised cell nucleus segmentation with active contours", Signal Processing, Vol. 71, no. 2, pp. 203-213, 1998.

[3] H. S. Wu, J. Barba, J. Gil, "A parametric fitting algorithm for segmentation of cell images", IEEE Transactions on Biomedical Engineering, Vol. 45, no. 3, pp. 400-407, 1998.

[4] M. H. Tsai, Y. K. Chan, Z. Z. Lin, S. F. Yang-Mao, P.-C. Huang, "Nucleus and cytoplast contour detector of cervical smear image", Pattern Recognition Letters, Vol. 29, pp. 1441–1453, 2008.

[5] E. Bak, K. Najarian, J. P. Brockway, "Efficient segmentation framework of cell images in noise environments", Proceedings of 26th Annual International Conference of the IEEE Engineering in Medicine and Biology, Vol. 1, pp. 1802-1805, 2004.

[6] N. Lassouaoui, L. Hamami, "Genetic algorithms and multifractal segmentation of cervical cell images", Proceedings of Seventh International Symposium on Signal Processing and its Applications, Vol. 2, pp. 1-4, 2003

[7] N. A. Mat Isa, "Automated edge detection technique for Pap smear images using moving K-means clustering and modified seed based region growing algorithm", International Journal of the Computer, the Internet and Management, Vol. 13, no. 3, pp. 45-59, 2005.

[8] M. E. Plissiti, C. Nikou, A. Charchanti, "Accurate localization of cell nuclei in pap smear images using gradient vector flow deformable models", Proceedings of 3rd International Conference on Bio-inspired Signals and Systems (BIOSIGNALS 2010), pp 284-289, 2010.

[9] A. N. Evans, "Morphological gradient operators for colour images", Proceedings of the IEEE International Conference on Image Processing (ICIP04), Vol. 5, pp. 3089-3092, 2004.

[10] C. Xu and J. Prince. "Snakes, shapes and gradient vector flow", IEEE Transactions on Image Processing, Vol. 7, No 3, pp 359-369, 1998.

[11] N. Otsu, "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, no. 1, pp. 62-66, 1979.

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