

# COMBINING SHAPE, TEXTURE AND INTENSITY FEATURES FOR CELL NUCLEI EXTRACTION IN PAP SMEAR IMAGES

Marina E. Plissiti<sup>1</sup>, Christophoros Nikou<sup>1</sup> and Antonia Charchanti<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Ioannina, Ioannina, Greece.

<sup>2</sup>Department of Anatomy-Histology and Embryology, Medical School, University of Ioannina, Ioannina, Greece.

*Abstract* — In this work, we present an automated method for the detection and boundary determination of cells nuclei in conventional Pap stained cervical smear images. The detection of the candidate nuclei areas is based on a morphological image reconstruction process and the segmentation of the nuclei boundaries is accomplished with the application of the watershed transform in the morphological color gradient image, using the nuclei markers extracted in the detection step. For the elimination of false positive findings, salient features characterizing the shape, the texture and the image intensity are extracted from the candidate nuclei regions and a classification step is performed to determine the true nuclei. We have examined the performance of two unsupervised (K-means, spectral clustering) and a supervised (Support Vector Machines, SVM) classification technique, employing discriminative features which were selected with a feature selection scheme based on the minimal-Redundancy – Maximal-Relevance criterion. The proposed method was evaluated on a data set of 90 Pap smear images containing 10248 recognized cell nuclei. Comparisons with the segmentation results of a gradient vector flow deformable (GVF) model and a region based active contour model (ACM) are performed, which indicate that the proposed method produces more accurate nuclei boundaries that are closer to the ground truth.

**Keywords:** Cell nuclei segmentation, Pap smear images, morphological reconstruction, watersheds, feature selection, clustering.

## I. INTRODUCTION

For over 30 years, the most effective and widespread screening test for cervical cancer is the Papanicolaou (Pap) test [1]. This technique provides a staining procedure of cervical cells, which results in the identification of the abnormalities in the cervix. The cervical cells are sampled and smeared onto a glass slide and the characterization of the slide (as normal or abnormal) is accomplished through the careful microscopical examination of the slide by an expert cytopathologist. Nowadays in developed

countries, the extensive use of the Pap test has significantly reduced the incidence and mortality of invasive cervical cancer.

Although the high diagnostic value of this test, the Pap smear images present certain limitations, which come from the fact that the conventional smear exhibits uneven layering, crowding and overlapping of cells. In addition, the staining introduces variances in illumination and dye concentration which result in inhomogeneities of the intensity of the cells. Therefore, the visual interpretation of these images is time-consuming and requires high level experience by the observer. For these reasons, in the last years many efforts have been made by several researchers in order to contribute to the automated analysis of such images.

The correct characterization of Pap smear slides and the derivation of conclusions for the contents of the Pap smear in a high degree depend on the general appearance of the cells nuclei. This is based on the fact that the nucleus is an important structural part of the cell which exhibits significant changes when a 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 in Pap smear images. Thus, the prerequisite for any further processing of Pap smear images is the accurate determination of the cell nuclei area. However, the exact nuclei locations in the image are not clearly defined in many cases, mainly due to cell overlapping in combination with the existence of many artifacts, and the nuclei boundaries are quite ambiguous. For this reason, two open problems pose a challenge for every method proposed for the automated analysis of Pap smear images: the exact detection of nuclei locations and the accurate determination of nuclei boundaries.

Some of the methods proposed in the literature deal only with one aspect of the problem, which is the segmentation of the cell nucleus and cytoplasm boundaries. The images that are used as test set, are presegmented from the original Pap smear images and they contain only one cell and consequently one single nucleus. Several image processing methods are proposed in this scope, such as active contours [2], template fitting [3] and edge detectors [4, 5, 6]. These methods exhibit remarkable performance in the segmentation of the structural parts of the cell. However the direct application of these methods in original Pap smear images, which may contain a large number of cells, cell overlapping and image artifacts is not appropriate, as they are focused on the recognition of the boundaries of the nucleus and the cytoplasm in images which contain only one single cell.

More sophisticated approaches to the automated analysis of Pap smear images are the methods which are applied on images containing a large number of cells in cell clusters, which are clearly more complicated. These methods manage to exclude the background of the image and to recognize the locations and the boundaries of the cells. Several approaches have been proposed, such as deformable templates [7], genetic algorithms [8], region growing with moving K-means [9] and pixel classification schemes [10]. Although these methods present promising results, their evaluation is restricted in a small data set of images and the performance criterion that is used is visual inspection, from which no reliable results about the general behavior of these methods can be obtained.

Methods based on watersheds for the analysis of Pap stained images have also been proposed in the literature. In [11], images containing one single nucleus of a Pap stained squamous epithelial cell are oversegmented with the watershed transform in order to define the differently stained subareas of the nucleus. Furthermore, in [12] watersheds are used for the detection of isolated cells

in low resolution images. However, in both methods, the problem of the detection of the accurate nuclei boundaries has not been resolved. Furthermore, Lezoray *et al.* [13] proposed a method for the determination of nuclei boundaries in Pap stained serous cytologies using color watersheds, which requires the cooperation of pixel classification schemes for the extraction of the nuclei markers. Nevertheless, the rough assumption used by the pixel classification schemes that the pixels of the entire image are distributed in two discrete classes, such as nuclei pixels and other pixels, is not appropriate for Pap smear images, which exhibit great complexity and the separation of all the pixels of the image in only two classes would produce noisy results.

In our work, we propose a two-stage fully automated method for the accurate determination of the nuclei boundaries in Pap smear images, which may contain both isolated cells and cell clusters. More specifically, in the first step, nuclei markers are detected with a procedure based on morphological reconstruction for the extraction of the areas of regional minima in the image, which usually correspond to nuclei locations [14]. The centroids of the areas of the regional minima are considered as markers in the watershed transform for the extraction of the nuclei boundaries. The morphological color gradient image is used for the flooding process, in order to retain the color information of the image.

In the second stage, we extend the segmentation of nuclei boundaries with the determination of meaningful features of the detected areas, which contribute to the identification of the true nuclei in Pap smear images. It must be noted that several methods [15, 16] propose a number of cell features for the characterization of a cell as normal or abnormal. However, they involve images containing one single cell. Since our images contain overlapped cells and cell clusters, our aim is to identify the nuclei areas and to separate the results of the segmentation in two categories: the true nuclei and other findings. Therefore, from the extracted boundaries, features describing the shape and the texture of each segmented regions are calculated. In addition we also integrate texture features and intensity disparity features of the neighborhood of each detected area. The latter evince to be some of the most discriminative features by a feature selection step based on minimal-Redundancy – Maximal-Relevance (mRMR) criterion [17]. It must be noted that in our experiments we have estimated the mRMR feature rank with two different approaches, namely using the entire image data set and the “leave-one out” strategy, as it is explained in more details in the following paragraphs.

A classification step is then performed for the reduction of unwanted findings. In this framework, the performance of two unsupervised (K-means and the spectral clustering) and one supervised (Support Vector Machine, SVM) classification schemes were examined. Our method was evaluated not only for the correct identification of cells nuclei locations but also for the accurate determination of nuclei boundaries with the boundaries obtained using the Gradient Vector Flow (GVF) deformable model [18] and a region based active contour model (ACM) [19] in terms of the Hausdorff distance from the ground truth. The method was evaluated using a large data set of 90 Pap smear images containing 10248 recognized cell nuclei, and the results indicate that the proposed method demonstrates high performance in both detection and segmentation of nuclei boundaries.

## II. METHOD

### A. Detection of the nuclei markers

The first step of the proposed method is the detection of the nuclei markers in each image. This is accomplished following a two

stage procedure, which includes the image preprocessing and the estimation of candidate nuclei centroids. It must be noted that the nuclei markers are obtained automatically in both isolated cells and cell clusters in the image.

### 1) Preprocessing

The preprocessing step is necessary for the definition of the regions of interest in the image which are occupied by cells. In this step, a binary mask is extracted with the cell areas highlighted. For this purpose, the initial image is firstly enhanced with the application of the contrast limited adaptive histogram equalization [20] in all color components, in order to obtain an image with the cells area more pronounced. Then, a global threshold is extracted using the standard method proposed by Otsu [21] in each color component, and the extracted binary masks are added with a logical OR operator. In the final mask, the detected areas are extended with a morphological dilation, and then all connected components with an area smaller than the area of an isolated cell are removed. This is required for the elimination of small objects that usually correspond to image artifacts, which may interfere in the classification step of the method.

### 2) Estimation of candidate nuclei centroids

After the preprocessing step, an image with the background extracted is obtained. Given the fact that the nuclei are darker than the surrounding cytoplasm, we search for intensity valleys in the detected areas. This is accomplished with the morphological reconstruction of the image and the detection of regional minima.

More specifically, the  $h$ -minima transform [22] is applied in the red, green and blue channels of the original image. A mask image is produced from the original image, with the subtraction of a constant value  $h$  from its pixels. Using this mask, a morphological reconstruction [23] of the original image is performed and this results in the formation of homogenous minima in the image. Afterwards, the detection of regional minima is obtained with the application of the non regional maxima suppression [24] in the complement of the derived image and the boundaries of the intensity valleys are estimated.

In each detected regional minima area, the coordinates of the centroid  $r_c$  is defined as:

$$r_c = (\bar{x}, \bar{y}) = \frac{1}{N} \sum_{i=1}^N (x_i, y_i), \quad (1)$$

where  $N$  is the number of pixels consisting the boundary of the regional minimum, and  $x_i, y_i$  are the coordinates of the pixel  $i$  of the boundary. These centroids indicate the probable position of the nuclei in the image. However, due to the inhomogeneity in dye concentration in the area of the nucleus, the detection of multiple centroids within the area of a single nucleus is possible. For the determination of a unique centroid in each nucleus, the elimination of two (or more) centroids in a regional minimum of a radius that it is smaller than the mean radius of a normal nucleus is required. This is accomplished with the application of the distance dependent rule, which is described as:

repeat

$$\forall p=(x,y) \in R_c$$

$$\text{if exists } q = \{(x_q, y_q) | D(p, q) \leq T\}$$

$$\text{select } r = \{p, q | \min\{I(p), I(q)\}\}$$

update  $R_c$

until no change in  $R_c$ .

In this rule,  $R_c$  is the set of all centroids,  $D$  is the euclidean distance between two points,  $T$  is the threshold on the minimum radius and  $I(p)$  is the intensity of the image at the point  $p$ . The detected nuclei markers in an image with cell overlapping are depicted in Fig. 1(b). The next step of our method is the construction of the morphological color gradient image, which exploits the color information of the original image in order to enhance the nuclei boundaries.

### B. Morphological Color Gradient Image

For the application of the watersheds, an image containing pronounced nuclei boundaries is required. Given the fact that most of the nuclei usually have ellipse-like shape, with the intensity of the pixels inside the nucleus area lower than those lying outside, high gradient of the image across the nuclei boundaries is expected. However, the extensive variances in nuclei intensity which are present due to the staining procedure result in gradient values of nucleus/cytoplasm borders that fluctuate in a wide range. For this reason, the use of a threshold after the application of edge detectors in order to determine the nuclei edges in the image would produce noisy results, because low thresholds would result in the detection of too many false edges, while high values would result in the loss of some true nuclei boundaries (Fig. 2). Therefore, we construct a gradient image using the color morphological gradient [25], in order to exploit the color information of the image for the estimation of the nuclei borders.

In general, the morphological gradient of a grayscale image  $f$  is defined as:

$$\nabla(f) = \delta_g(f) - \varepsilon_g(f), \quad (3)$$

where  $\delta_g(f)$  and  $\varepsilon_g(f)$  is the grayscale dilation and grayscale erosion for a structuring element  $g$  respectively. Alternatively, the morphological gradient can be expressed as:

$$\begin{aligned} \nabla(f) &= \max_{x \in g} \{f(x)\} - \min_{x \in g} \{f(x)\} \\ &= \max(|f(x) - f(y)|) \quad \forall x, y \in g \end{aligned} \quad (4)$$

which is the maximum absolute intensity difference between two pixels in the area of the structuring element. For color images with pixels denoted as three dimensional vectors the color morphological gradient (CMG) can be expressed as:

$$\text{CMG} = \max_{i, j \in G} \left\{ \|x_i - x_j\|_p \right\} \quad (5)$$

where  $x_i, x_j$  are pixels in the structuring element  $G$ . In our experiments we compute the second norm ( $p=2$ ) and the structuring element that is used is a  $3 \times 3$  flat structuring element. The color morphological gradient of a representative Pap smear image is depicted in Fig. 1(c).

### C. The Watershed Transform

The concept of watersheds [26] in image processing is based on considering an image in three dimensional space, with two spatial coordinates versus intensity. The value of the intensity is assumed to be the elevation information. In terms of this topographic representation of the image, the pixels are divided into three categories: pixels of regional minima, pixels of catchment basins and pixels of watershed lines, which separate neighboring catchment basins and consequently they separate different characteristic parts of the image.

For the detection of the watershed lines in an image  $I$  with regional minima  $M_1, M_2, \dots, M_R$ , a flooding process is performed in integer flood increments from  $n_0 = \min(I) + 1$  to  $n_{\max} = \max(I) + 1$ . Let  $C(M_i), i = 1, \dots, R$  be the sets of points in the catchment basin corresponding to the regional minimum  $M_i$  and let  $C[n]$  be the union of the flooded catchment basins at stage  $n$  of the flooding process. The set of the image points with intensity value lower than  $n$  is defined as  $T[n] = \{p | I(p) < n\}$ . The above sets of points are initialized as  $C[\min(I) + 1] = T[\min(I) + 1]$ . In the next steps of the algorithm, the set  $C[n]$  is sequentially derived from  $C[n-1]$  as follows:

Let  $Q$  be the set of the connected components in  $T[n]$ . Then for each connected component  $q \in Q$  the intersection ( $\lambda$ ) with the set  $C[n-1]$  is calculated as  $\lambda = q \cap C[n-1]$ . Depending on the value of  $\lambda$  there are three possibilities:

- a) If  $\lambda$  is empty then a new minimum is present and the connected component  $q$  is added into  $C[n-1]$ , thus  $C[n] = C[n-1] \cup q$ .
- b) If  $\lambda$  contains one connected component of  $C[n-1]$  then  $q$  belongs to an existing catchment basin of a regional minimum and consequently  $C[n] = C[n-1] \cup q$ .
- c) If  $\lambda$  contains more than one connected component of  $C[n-1]$  this means that  $q$  partially belongs to different catchment basins and the next step of flooding would cause the water level in these catchment basins to merge. For this reason, a watershed line must be constructed to prevent the overflow between these catchment basins.

The application of the watershed transform in this form usually results in oversegmentation of the image, because of the presence of artifacts and noise. To avoid this undesirable effect, the watersheds are applied in edge images with markers, which are connected components belonging to specific regions of interest in the image and they are used as starting points of the flooding process.

In the proposed method the nuclei markers are determined automatically with the morphology based processing scheme, as it is described above. Furthermore, we perform the distance transform in the binary mask obtained in the preprocessing step, in order to construct the cytoplasm markers. The result of the watershed transform in an image with nuclei markers is depicted in Fig. 1(d).

#### *D. Clustering of the candidate nuclei*

The determination of the watershed lines, usually results in the correct identification of the nuclei positions in the image. However, some false positive areas are also detected, due to the existence of a regional minimum. This is a consequence of the detection of the nuclei markers step, which produces some centroids of regional minima that do not indicate the existence of nuclei (Fig. 3). Therefore, the elimination of these areas is necessary and a clustering step is performed for the separation of the detected areas into two classes: the true nuclei class and the rest of the findings. Thus, for every detected area a vector of features is determined, which will be used as input to the clustering algorithms.

##### *1) Feature extraction*

The efficient separation of the true nuclei regions from the total segmented regions requires the generation of meaningful features of very good discriminative ability. Having found the areas of the nuclei enclosed by the detected boundaries, features concerning the shape, the texture and the intensity of the detected regions can be easily determined. However, the restriction of the calculation of these features only for the area enclosed by the detected boundaries is not sufficient because regions of regional minima not corresponding to true nuclei may also have similar features. In this step it is expedient to take advantage of the fact that the nuclei are darker than the surrounding cytoplasm and the detected nuclei regions would present significant differences from their neighborhood. Moreover, the detected regions that do not belong to nuclei were probably detected due to the existence of shallow minima in the intensity of the area of the cytoplasm in the image, and they are more likely to present similar features values from their neighborhood (Fig. 4).

For this reason, we propose the calculation of features also for the neighborhood of the detected areas, which is defined in terms of the bounding box of these areas (Fig. 5). More specifically, for each detected area  $A$ , the bounding box  $B$  is calculated as the maximum rectangle that contains the detected region, and the neighborhood  $Ngh$  is determined as the complement  $A^c$  in  $B$ , that is  $Ngh = A^c \cap B$ . In our work, for the construction of the feature set, the pixels within the detected region, the pixels of the neighborhood and the pixels of the bounding box are taken into account. Three categories of features are developed: shape, textural and intensity disparity features.

##### *i) Shape Features*

The detected boundaries for the nuclei are expected to present an ellipse-like shape and several features to describe this characteristic are chosen. More specifically, six features extracted from the shape of the detected region boundary are calculated, that is the Circularity, the Eccentricity, the Major and the Minor Axis Length, the Equivalent Diameter of a circle with the same area as the region, and the Perimeter of the detected region. The Major Axis Length, the Minor Axis Length and the Eccentricity are defined in terms of an ellipse that has the same central second moments as the region. The shape features are presented in Table I.

### *ii) Textural Features*

The texture analysis of the detected regions is based on the statistical properties of the intensity histogram in the three color components and the calculation of some texture descriptors such as the local binary patterns (LBP, see Appendix A) [16, 27]. Thus, for every segmented region we have calculated the Third Moment, the Uniformity, the Entropy and the Smoothness of the intensity histogram for the three predefined regions ( $A, B, Ngh$ ). Moreover, the normalized uniform rotation-invariant LBP occurrence histogram was calculated for the bounding box ( $B$ ) of the segmented regions, using LBP of two different neighborhood topologies: a circle of unit radius and a hyperbola with semi-major and semi-minor axis lengths equal to one (fig. 6). In both topologies, the number of equally spaced pixels was  $P=8$  (see Appendix A for more details and [16] and [27] for a more in depth explanation of these features). The mean and the standard deviation of each histogram were used as features. All the textural features are calculated for all three color channels and they are summarized in Table II.

### *iii) Intensity Disparity Features*

The feature that characterizes the intensity of each region is the average of the intensity value of all the pixels of the region. However, as it is observed, the average intensity of the nuclei varies in a wide range and may coincide with regions of cell overlapping in the image. An equivalent intensity feature that pronounces the disparity of the detected region and its neighborhood is the difference of the average intensity between those regions (Table III). We expect high values for this feature when it refers to nuclei regions, as the nuclei area is darker than the surrounding cytoplasm. Three values of this feature were calculated independently for the red, green and blue component of the original image.

## *2) Feature selection*

For each detected region we have calculated in total 57 features. More specifically, 6 features concerning the shape of the region, 3 features concerning the intensity disparity of the detected areas and their neighborhood and finally, for the three color components,  $3 \times 4$  textural features for the enclosed area ( $A$ ),  $3 \times 4$  textural features for the neighborhood ( $Ngh$ ),  $3 \times 8$  textural features for the bounding box ( $B$ ) were calculated. However, the contribution of each feature is different in the categorization of the data. For the selection of the most discriminative features, a feature selection technique is employed which is based on the Minimal-Redundancy-Maximal-Relevance (mRMR) criterion [17]. More specifically, given a data set of  $N$  samples of  $M$  features  $X = \{x_i^j, i = 1, \dots, M, j = 1, \dots, N\}$ , and the target classification variable  $c$ , the objective is to find from the  $M$  dimensional space  $R^M$  a subset of  $m$  features that characterizes  $c$  more efficiently.

The mRMR criterion combines both Max-Relevance ( $\max D$ ) and Min-Redundancy criteria ( $\min R$ ), which are defined respectively as [17]:



$$\max_{S \subset X} D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c), \quad (6)$$

$$\min_{S \subset X} R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j), \quad (7)$$

where  $S$  is the feature set and  $I(x; y)$  is the mutual information between two random variables, which is defined in terms of their marginal and joint probability density functions  $p(x)$ ,  $p(y)$  and  $p(x, y)$  as:

$$I(x; y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. \quad (8)$$

The mRMR criterion is then defined as:

$$\max_{S \subset X} (D(S, c) - R(S)). \quad (9)$$

The selection of features for the construction of the final set is obtained incrementally, that is if  $m-1$  features are already selected in the  $S_{m-1}$ , then the  $m^{\text{th}}$  selected feature will be the one that satisfies eq. (9). The optimal size of the features set depends on the specific classification algorithm that will be used.

Thus the features were ranked in a range beginning from the most powerful discriminative feature to the feature with the least discriminative power. It must be noted that for the calculation of the mutual information, each feature variable was discretized into three states at the positions  $\mu \pm std$  ( $\mu$  is the mean value and  $std$  is the standard deviation of the specific feature distribution). More specifically, it takes -1 if the feature value is less than  $\mu - std$ , 1 if the feature value is larger than  $\mu + std$  and 0 otherwise. This assumption is reliable when our features follow a unimodal-like distribution. This was verified by the construction of the histograms of each feature and some representative examples are depicted in Fig. 7. In Table IV the first 16 most discriminative features for all the segmentation techniques are presented.

### 3) Clustering algorithms

In our work, for comparison purposes, three clustering methods are employed for the separation of the detected areas in the true nuclei class and the other findings class: the K-means [28], the spectral clustering [29] (see Appendix B) and the Support Vector Machine (SVM) classifier with the radial basis function (RBF) kernel [30]. Given the fact that the K-means and the spectral clustering algorithms do not require any training, they are applied independently in each image. However, for the application of the SVM classification algorithm a training data set is constructed. In our experiments, we use the “leave one out” technique for the evaluation of the performance of the classifier. Thus, 21 slides were used as training set and the remaining slide was used as test set. This experiment was repeated 22 times, each time using a different slide as test set. The performance of the classification is calculated using the trained SVM classifier in the test set.

### III. RESULTS

#### 1) Study Group

We have collected 90 images from 22 different 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 of size 1536×2048 were stored in JPEG format. The total number of cell nuclei in the images, which were identified by an expert observer is 10248. In order to obtain the ground truth, the nuclei locations were manually identified.

#### 2) Numerical Evaluation

The presented method was bilaterally evaluated in order to estimate the performances of the clustering algorithms for the detection of the true nuclei in the images, and also the accuracy of the segmentation, in comparison with the ground truth (manually traced nuclei boundaries). Furthermore the method performance was compared with the corresponding performance of two different segmentation techniques, namely the GVF deformable model [18] and the ACM model [19] in terms of both classification and segmentation results. In the detected regions of both GVF and ACM segmentation techniques, the previously described features were determined.

For the classification performance, we have calculated the number of true positive ( $TP$ ), true negative ( $TN$ ), false positive ( $FP$ ) and false negative ( $FN$ ) findings in all images of our data set. Two widely used statistical measures for the performance of the classification are calculated:

1. The *sensitivity*, which measures the proportion of actual nuclei which are correctly identified as such and it is defined

$$\text{as: } \textit{sensitivity} = \frac{TP}{TP + FN} . \quad (10)$$

2. The *specificity* which measures the proportion of candidate centroids that are not nuclei and are correctly

$$\text{characterized as such by the classification techniques, and it is defined as: } \textit{specificity} = \frac{TN}{TN + FP} . \quad (11)$$

In addition, the segmentation performance was evaluated with the calculation of the Hausdorff Distance ( $D_{\text{Hausdorff}}$ ) between the manual traced boundary  $M$  and the boundary  $\Psi$  obtained from the segmentation procedure defined as:

$$D_{\text{Hausdorff}} = \max_{a \in M} \left\{ \min_{b \in \Psi} \{D(a, b)\} \right\} \quad (12)$$

where  $D$  is the Euclidean distance.

It must be noted that in the detection of the nuclei markers the method misses in total 147 true nuclei position which is a total loss rate of 1.01%. Thus in the following steps, the total number of true nuclei is reduced to 10101.

#### A) Classification

In our experiments we have tested several configurations of the classification process which involve both the calculation of the mRMR feature rank and the clustering algorithms (K-means, spectral clustering and SVM). For this reason we include the following experiments, which were executed for all the data sets obtained from the three segmentation algorithms (watersheds, GVF, level sets):

- a. The estimation of mRMR rank of feature was determined in two different ways:
  - The whole data set of patterns was used as input to the mRMR criterion (global mRMR) and a ranking was obtained which was then used in the classification algorithms.
  - The set of patterns was separated into 22 folds (each fold corresponds to a single slide). Then, 21 folds were used for training and the remaining fold was used for testing. From the training set, we obtained the mRMR rank (leave-one-out mRMR) of features and this rank was used in the classification algorithms applied to the image of the testing slide. This procedure was repeated 22 times, each time using a different slide as test set. By these means, we obtained 22 different feature ranks, which were assigned in the 22 folds (slides).

Therefore, all of the classification techniques (K-means, spectral clustering and SVM) were executed twice, using the above mRMR rankings (global mRMR and leave-one-out mRMR). For the selection of the ideal number of features, the performance of the classification techniques was estimated on the test set using a pattern of increasing dimension varying from 2 to 57 features. Starting from a pattern described by only 2 features, one feature was added incrementally until all of the 57 features are employed. In the second case described above, the selection of the feature that is added in the pattern is different for each test slide (and consequently for the images belonging to this slide) and it was determined by the corresponding mRMR rank (obtained using the other 21 slides as training set). In order to evaluate the importance of each feature, the mean position and its standard deviation in a feature histogram was calculated (Fig. 8).

- b. The estimation of the best value for parameter  $\sigma$  in spectral clustering was also obtained using a leave-one-out strategy. The set of patterns was separated into 22 folds (each fold corresponds to a single slide), with 21 folds were used for training and the remaining fold was used for testing. Several experiments with different values for  $\sigma$  were performed in the training set, using patterns containing all of the features (the dimension of each pattern was 57). Then, we selected the value of  $\sigma$  that exhibited the best performance in the training set. This value was used for the application of spectral clustering in the images of the test set. This procedure was repeated 22 times, each time using a different training and test set.
- c. The values of the parameters of the SVM classifier ( $\gamma$  and  $C$  for the RBF kernel) were obtained by constructing two different data sets, each one containing half of the slides (11 slides were randomly selected for the training set and the remaining were used as test set). We performed several experiments with different pairs of values for  $\gamma$  and  $C$  ( $(C, \gamma) \in [0.01, 0.125, 0.25, 0.5, 1, 2, 4, 8]$ ), while the SVM classifier was trained with the training set of patterns containing 57 features. Afterwards the performance of the classifier was estimated with the test set. The values for  $\gamma$  and  $C$  were selected as

those which exhibit the best performance of the SVM classifier in the test set and they were  $\gamma=0.01$  for all the segmentation methods and  $C=2$  for the GVF segmentation and  $C=4$  for the watershed and the ACM segmentation.

The number of features that results in the best classification performance depends on the specific classification algorithm. When a performance criterion is maximized for a specific number of features, then this subset of features is selected. In our work, the performance criterion that the clustering algorithm should maximize is the harmonic mean (HM) of the *sensitivity* and the *specificity* defined as:

$$HM = \frac{2 \times sensitivity \times specificity}{sensitivity + specificity} \quad (12)$$

In Fig. 9, the values of HM criterion versus the number of features are depicted for the ACM, GVF and the watershed segmentation for the K-means algorithm. Similar experiments were performed for the definition of the best feature subset using the spectral clustering algorithm (Fig. 10) and the SVM classifier (Fig. 11). The performance of the SVM classifier for the watershed and the GVF segmentation increases as more features are used, and reaches the maximum performance at 57 features. For the ACM segmentation, the SVM classifier reaches the maximum performance at 26 features. In all cases as it can be observed, the HM measure for the watershed segmentation is higher than the other two segmentation techniques.

More specifically, the best results in terms of the HM for the all the segmentation schemes using the global and the leave-one-out mRMR rank are presented in Table V. As we can see, the best results were obtained with the K-means clustering algorithm using patterns obtained from the watershed segmentation. The SVM classifier is preferable for the ACM and GVF segmentations, as it produced higher performances than the K-means and the spectral clustering. Furthermore, in most of the cases, the use of leave-one-out mRMR feature rank produces better results in comparison with the use of the global mRMR rank. It must be noted that for comparison purposes, the performance of the SVM classifier was selected for 26 features for all segmentation techniques.

### B) Segmentation

In order to evaluate the performance of the segmentation method, the obtained nuclei boundaries were compared with the corresponding resulted nuclei boundaries of the GVF deformable model and the ACM model and also with the manually traced boundaries. It must be noted that for the application of the GVF deformable models, an initial approximation of every nucleus boundary is required. For this reason, we search for some points in the neighborhood of each detected centroid, which are likely lying in the nucleus circumference [31]. In the morphological color gradient image, having as starting points the candidate nuclei centroids we construct a circular searching grid with 8 radial profiles consisted of 8 points each and centered at the location of each candidate nucleus centroid. In each radial profile we choose the pixel with the highest intensity (non maximum suppression) and the initial approximation of the nuclei boundaries is obtained with the convex hull of the circumferential points found in the this step. The values for the weighting parameters of the GVF deformable model are fixed for all the images and they are set to be  $a = 0.9$  for the tension,  $\beta = 1.5$  for the rigidity and  $\gamma = 3$  for the image force.

In a similar way, the ACM model was also applied to the same images. More specifically, having found the nuclei markers, we apply the ACM model, as it is described in [19] in the  $21 \times 21$  image window centered at each marker. The model was initialized as

a rectangle in the middle of the selected neighborhood and it was applied in the morphological color gradient image with  $\alpha=20$ , where  $\alpha$  is the balloon force which controls the contour shrinking or expanding.

Several examples of the segmentation results are depicted in Fig. 12. The Hausdorff Distance for the ground truth and the watershed segmentation was estimated as  $1.71\pm 0.54$  (*mean $\pm$ std*). The corresponding distance for the GVF and ACM segmentation is  $2.65\pm 3.23$  and  $2.48\pm 2.30$  respectively. This implies that the watershed segmentation is closer to the manually traced nuclei boundaries, and as a result it is more accurate than GVF and the ACM segmentation. Furthermore, the ACM segmentation is more performing than the GVF segmentation, as it exhibits lower Hausdorff Distance. In the next paragraph, some reasons of failure for the GVF and ACM segmentations are discussed.

#### IV. DISCUSSION

The proposed method for the segmentation of the cell nuclei in Pap smear images is fully automated and it can be applied directly in any conventional Pap stained cervical smear images, in order to produce accurate nuclei boundaries. It consists of five steps: the preprocessing, the estimation of the candidate nuclei centroids, the application of the watershed transform, the feature extraction and the classification step. The method was developed in Matlab using a dual core PC with a 2.0 GHz processor and 3GB of RAM. The execution time for each step of the method depends on several factors, such as the proportion of the image characterized as background in the preprocessing step, the number of the candidate nuclei centroids in each image, the classification algorithm and the number of features in each pattern. An indicative execution time for the segmentation of the images (steps one to four of our method) is 2-5 min. The mean execution time of K-means in an image using 16 features is less than a second, while the corresponding time for the spectral clustering algorithm is 5-6 seconds. Finally, the mean execution time for the training of the SVM classifier using 21 slides and the evaluation of the performance in the test set (one slide) varies from 2 to 4.5 minutes.

The parameters used in the several steps of the segmentation method were determined after careful examination of the images by an expert cytopathologist in combination with the results of several tests. Thus, in the preprocessing step, for the elimination of small objects that do not correspond to nuclei positions, we used as a threshold of 500 for the object area, which is sufficient for the rejection of small image artifacts, while preserving the isolated cells in the image. Furthermore, for the extraction of the nuclei markers in the images with the  $h$ -minima transform, we use a threshold value of  $h=15$  for the extraction of the regional minima. In addition, in the distance dependent rule, for each detected centroid we calculated the minimum Euclidean distance from the neighboring centroids and we used a threshold of 8 for the minimum radius of the nucleus neighborhood.

However, as it was mentioned before, this step misses some of the true nuclei positions. This is mainly due to the faintly staining and the uneven layering of some cells. In the first case, the cells are undistinguished from the background and as a consequence, the nuclei of these cells are considered as isolated objects in the image background and they are rejected as image artifacts. In the second case, the intensity of the nucleus does not well differentiate from the cytoplasm intensity and no regional minimum is detected in the nucleus position.

The nuclei markers obtained in the previous step are used in the application of the watershed transform. The importance of this step is crucial, as it prevents from the oversegmentation that would be produced by the application of the watershed transform in the images without markers. Hence, using the detected cytoplasm markers, the flooding process starts from a position in the catchment basins of the nuclei area and finally converges to the actual boundaries of the true nuclei. Furthermore, the problem of the detection of false positive detected centroids is effectively resolved in the classification step.

The feature selection using the mRMR criterion produces different feature ranks for the three segmentation techniques, as it can be observed in Table IV. This is a consequence of the differences between the segmented regions provided by each method (ACM, GVF, watersheds) necessitating different features for its representation. As it can be observed by the feature ranking, we can conclude that the discriminative ability of some features is equally important for all the segmentation techniques, as seven of them were selected by all of the segmentation techniques in the first 14 positions. These features are highlighted in bold face fonts in Table IV. Furthermore, from Fig. 8 we can observe that in general, the standard deviation of the features selected by the leave-one-out mRMR is rather insignificant for the first 10 and the last 20 positions in the mRMR rank, which indicates that from the entire data set of features, the most discriminative and the least discriminative features are the same for every fold (slide) of our image data set.

As it was verified by the results, the watershed segmentation is more accurate than the GVF and the ACM segmentation. For both these segmentation techniques, the main reason of failure is that their behavior highly depends on the values of their parameters. Furthermore, the existence of a high gradient value in a small distance of the detected nucleus and the inhomogeneities on the nuclei intensity affect the performance of these techniques. Some examples of these cases are depicted in Fig. 13. In Fig. 13(a) as the gradient in the border of the nucleus/cytoplasm is weak, the shape of the GVF deformable model is mainly determined by its internal forces, which enforce it to be of a relatively small length and smooth. Furthermore, in Fig. 13(b) the existence of intensity variations in the area of the nucleus attracts the points of the GVF deformable model, which converges to a position far from the actual nucleus boundary. For the same images, the ACM model also fails to accurately determine the nucleus borders. Finally, in Fig. 13(c) both the GVF and ACM model are attracted by the points of high image gradient, which do not correspond to the boundary of the detected nucleus. In all these cases the GVF and ACM model do not succeed in detecting the accurate nucleus boundary. In contrast, as it is observed, the watersheds overcome these limitations and produce nuclei boundaries that are closer to the ground truth.

The accurate determination of the nuclei boundaries leads to the calculation of more accurate features, which improve the performance of the clustering algorithms. This is the reason why the use of features extracted with the watershed segmentation present better classification performance than the corresponding features extracted from the GVF and the ACM segmentation. Furthermore, for the determination of a feature set we exploit the fact that the true nuclei area presents significant variations with respect to its neighborhood and the calculation of neighborhood features would result in the effective discrimination of the true nuclei areas and the false positive areas. This is also confirmed by the use of mRMR criterion (Table IV), which indicates that for

the feature set obtained with all the segmentation techniques (Watersheds, GVF, ACM), at least 7 out of 10 most discriminative features concern the outer area (bounding box  $B$  and neighborhood  $Ngh$ ) of the detected boundaries.

Traditionally, immediate fixation and staining of the cellular sample on the slide with 70% ethyl alcohol and Papanicolaou stain have been established as the professional standard. This fixation and staining combination results in a cellular sample, that, not only has well-defined and tinted morphological features, but also its transparency allows for microscopic visualization of nuclear and cytoplasmic boundaries through multiple layers of epithelial cells. In our work, we use 90 conventionally stained Pap smear images, which exhibit several differences in colorization (e.g. the blue color can vary from deep blue to light blue). Although we have not included any process of color correction and detection of improper staining, the method provides accurate results when it is applied to the images of our data set. However, an issue that is under research is the separation of clustered nuclei, since the method in its current form indicates the existence of one nucleus in the specific location. Furthermore, the recognition of abnormal nuclei in Pap smear images is another issue that we consider as future work.

## V. CONCLUSION

The identification of the cervical cell nuclei areas is a prerequisite for the derivation of diagnostic conclusions and the characterization of the contents in the Pap smear images. The automated detection and segmentation of the nuclei boundaries in these images is a challenging issue, as these images present several limitations. In this work, we have effectively overcome the problem of the detection of the nuclei locations and we have developed a fully automated method for the segmentation of cell nuclei in Pap smear images. Moreover, we propose the determination of a meaningful feature set for the detected areas, which results in the efficient discrimination of the true nuclei class by the clustering algorithms. As it is verified by the results, the method produces more accurate nuclei boundaries which are closer to the ground truth, compared to the GVF deformable model and the ACM segmentation method. The main advantage of the proposed method is that it can be applied directly in Pap smear images obtained by an optical microscope, without any observer interference, for the accurate automated identification of the cell nuclei boundaries.

## ACKNOWLEDGMENT

The authors would like to thank Olga Krikoni for providing the Pap smear slides for the construction of the image data set.

## APPENDIX A

According to [27], the texture  $T$  in a local neighborhood of a monochrome image is the joint distribution of the gray levels of  $P$  ( $P>1$ ) image pixels:

$$T = t(g_c, g_0, \dots, g_{p-1}),$$

where gray value  $g_c$  corresponds to the gray value of the center pixel of the neighborhood and  $g_p$  ( $p = 0, \dots, P-1$ ) correspond to the gray values of  $P$  equally spaced pixels on a loci of points (usually a circle with radius  $R$  ( $R > 0$ )). By subtracting the gray value of the center pixel  $g_c$  from the gray value of the neighborhood pixels, we obtain an equivalent form of the texture, that is:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (A1).$$

If we assume that the differences  $g_p - g_c$  are independent of  $g_c$  the (A1) can be factorized as:

$$T \approx t(g_c) t(g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (A2)$$

and since the distribution  $t(g_c)$  describes the overall luminance of the image, it does not provide useful information for texture analysis, leading to a simplified form of (A2)

$$T \approx t(g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (A3)$$

which is a highly discriminative texture operator, as it records the occurrences of various patterns in the neighborhood of each pixel in a  $P$ -dimensional histogram. The invariance with respect to the scaling of the gray scale is achieved by considering just the signs of the differences  $g_p - g_c$  and not their exact value:

$$T \approx t(s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_{P-1} - g_c)) \quad (A4)$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

By assigning a binomial factor  $2^p$  for each sign  $s(g_p - g_c)$ , the (A4) is transformed into a unique  $LBP_{P,R}$  number that characterizes the spatial structure of the local image texture:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p .$$

It is observed that certain local binary patterns are fundamental properties of texture, and they are characterized as ‘‘uniform’’. The uniformity measure  $U(\text{pattern})$  corresponds to the number of spatial transitions (bitwise 0/1 changes) in the pattern. In general, the operator for grayscale texture description using rotation invariant uniform patterns introduced by [27] is defined as:

$$LBP_{P,R}^{riu} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{if } U(LBP_{P,R}) \leq 2 \\ P+1 & \text{otherwise} \end{cases}$$

where  $U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)|$ .

## APPENDIX B

### *Spectral clustering algorithm*

Given a set of vectors  $(x_1, x_2, \dots, x_N)$ ,  $x_k \in R^p$  and the number  $c$  of desired clusters to be separated, the spectral clustering algorithm performs the following steps:



1. Define the affinity matrix  $A^{N \times N}$  as  $A_{i,j} = \exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)$ .
2. Define the diagonal matrix  $D^{N \times N}$  as  $D_{ii} = \sum_j A_{ij}$ .
3. Define the matrix  $L = D^{-1/2} A D^{-1/2}$ .
4. Define the  $c$  largest eigenvalues  $\lambda_i, i = 1, \dots, c$  of  $L$  and the corresponding eigenvectors  $y_i, i = 1, \dots, c$ .
5. Form the matrix  $Y$  which has as columns the eigenvectors  $y_i$ .
6. Normalize each row of  $Y$  to have unit length.
7. Treat each row of  $Y$  as a point in  $R^c$  and cluster them into  $c$  clusters via K-means.
8. Assign the original points  $x_i$  to cluster  $j$  if and only if row  $i$  of the matrix  $Y$  was assigned to cluster  $j$ .

#### REFERENCES

- [1] G. N. Papanicolaou, A new procedure for staining vaginal smears, *Science*, 95 (2469), 1942, 438-439.
- [2] P. Bamford, B. Lovell, Unsupervised cell nucleus segmentation with active contours, *Signal Process.*, 71 (2), 1998, 203-213.
- [3] H. S. Wu, J. Barba, J. Gil, A parametric fitting algorithm for segmentation of cell images, *IEEE Trans. Biomed. Eng.*, 45(3), 1998, 400-407.
- [4] C. H. Lin, Y. K. Chan, C. C. Chen, Detection and segmentation of cervical cell cytoplasm and nucleus, *Int. J. Imaging Syst. Technol.*, 19(3), 2009, 260-270.
- [5] M. H. Tsai, Y. K. Chan, Z. Z. Lin, S. F. Yang-Mao, P. C. Huang, Nucleus and cytoplasm contour detector of cervical smear image, *Pattern Recognit. Lett.*, 29, 2008, 1441-1453.
- [6] S. F. Yang-Mao, Y. K. Chan, Y. P. Chu, Edge enhancement nucleus and cytoplasm contour detector of cervical smear images, *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, 38 (2), 2008, 353-366.
- [7] A. Garrido, N. Perez de la Blanca, Applying deformable templates for cell image segmentation, *Pattern Recognit.*, 33 (5), 2000, 821-832.
- [8] N. Lassouaoui, L. Hamami, Genetic algorithms and multifractal segmentation of cervical cell images, *Proceedings of 7th International Symposium on Signal Processing and its Applications*, 2, 2003, 1-4.
- [9] 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 Manage.*, 13 (3), 2005, 45-59.
- [10] 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*, 1, 2004, 1802-1805.
- [11] P. T. Jackway, Gradient watersheds in morphological scale space, *IEEE Trans. Image Process.*, 5, 1996, 913-921.
- [12] P. Bamford, B. Lovell, A water immersion algorithm for cytological image segmentation, *Proceedings of APRS Image segmentation workshop*, 1996, 75-79.
- [13] O. Lezoray, H. Cardot, Cooperation of color pixel classification schemes and color watershed: A study for microscopic images, *IEEE Trans. Image Process.*, 11 (7), 2002, 783-789.
- [14] M. E. Plissiti, E. E. Tripoliti, A. Charchanti, O. Krikoni, D. Fotiadis. Automated detection of cell nuclei in Pap stained smear images using fuzzy clustering, *Proceedings of 4th European Congress for Medical and Biomedical Engineering*, 2008, 637-641.
- [15] Y. Marinakis, M. Marinaki, G. Dounias, Particle swarm optimization for pap-smear diagnosis, *Expert Systems with Applications*, 35, 2008, 1645-1656.
- [16] L. Nanni, A. Lumini, S. Brahmam, Local binary patterns variants as texture descriptors for medical image analysis, *Artificial Intelligence in Medicine*, 49(2), 2010, 117-125.
- [17] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.*, 27 (8) 2005, 1226-1238.

- [18] C. Xu and J. Prince, Snakes, shapes and gradient vector flow, *IEEE Trans. Image Process.*, 7 (3), 1998, 359-369.
- [19] K. H. Zhang, L. Zhang, H. H. Song, W. Zhou, Active contours with selective local or global segmentation: A new formulation and level set method, *Image and Vision Computing*, 28(4), 2010, 668-676.
- [20] K. Zuiderveld, Contrast limited adaptive histogram equalization, *Graphics Gems IV*, 1994, 474-485.
- [21] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Trans. Syst. Man Cybern.*, 9 (1), 1979, 62-66.
- [22] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer-Verlag, New York, 1999.
- [23] L. Vincent, Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms, *IEEE Trans. Image Process.*, 2 (2), 1993, 176-201.
- [24] E. J. Breen, R. Jones, Attribute openings, thinings, and granulometries, *Comput. Vision Image Understanding*, 64 (3), 1996, 377-389.
- [25] A. N. Evans, Morphological gradient operators for colour images, *Proceedings of the IEEE International Conference on Image Processing (ICIP04)*, 5, 2004, 3089-3092.
- [26] R. C. Gonzalez, R. E. Woods, *Digital image processing*, second ed., Prentice Hall, 2002.
- [27] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Transactions on pattern analysis and machine intelligence*, 24 (7), 2002, 971-987.
- [28] C. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
- [29] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in *Advances in Neural Information Processing Systems*, 14, 2002, 849-856.
- [30] N. Christianini, J. S. Taylor, *Support Vector Machines and other kernel-based methods*, Cambridge University Press, 2000.
- [31] 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, 284-289.

---

Table I

Shape Features

---

Minor Axis Length <sup>(1)</sup>	$K = \sqrt{\frac{2(u_{20} + u_{02} - \Delta)}{u_{11}}}$
Major Axis Length <sup>(1)</sup>	$L = \sqrt{\frac{2(u_{20} + u_{02} + \Delta)}{u_{11}}}$
Eccentricity	$E = 2 \frac{\sqrt{\left(\frac{L}{2}\right)^2 - \left(\frac{K}{2}\right)^2}}{L}$
Equivalent Diameter	$ED = \frac{4 \times Area}{\pi}$
Perimeter	$P = \text{number of boundary points}$
Circularity	$C = \frac{4\pi \times Area}{P^2}$

---

<sup>(1)</sup> The formulas for  $\Delta$  and the central moments  $u_{pq}$  of order  $p+q$  of the region  $s(x,y)$  are defined as:

$$\Delta = \sqrt{4u_{11}^2 + (u_{20} - u_{02})^2},$$

$$u_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q, \text{ where } \bar{x} \text{ and } \bar{y} \text{ are the coordinates of the centroid of the region.}$$

---

Table II

Texture Features

---

Third Moment <sup>(2)</sup>	$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$
Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$
Entropy	$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$
Smoothness	$R = 1 - \frac{1}{1 + s^2}$ , where $s = \sqrt{\sum_{i=0}^{L-1} (z_i - m)^2 p(z_i)}$
Mean Histogram $LBP_{circle}^{riu2}$	See Appendix A
Std Histogram $LBP_{circle}^{riu2}$	See Appendix A
Mean Histogram $LBP_{hyperbola}^{riu2}$	See Appendix A
Std Histogram $LBP_{hyperbola}^{riu2}$	See Appendix A

---

<sup>(2)</sup> Given that  $z_i$  is the intensity value  $i$  and  $p(z)$  is the histogram of the intensity levels in a region with  $L$  possible intensity

levels, then the average intensity of the region is calculated as  $m = \sum_{i=0}^{L-1} z_i p(z_i)$ .

---

Table III

Intensity Disparity Features

---

Foreground-Background contrast in red <sup>(3)</sup>	$dR = m_{RED}^{Ngh} - m_{RED}^A$
Foreground-Background contrast green <sup>(3)</sup>	$dG = m_{GREEN}^{Ngh} - m_{GREEN}^A$
Foreground-Background contrast in blue <sup>(3)</sup>	$dB = m_{BLUE}^{Ngh} - m_{BLUE}^A$

---

<sup>(3)</sup>  $m_{color}^{region}$  is the average intensity value of an image region in a specific color component. The RGB color space is used in our experiments and the regions of the image that are considered are the enclosed boundary area  $A$  and its neighborhood  $Ngh = A^c \cup B$ , where  $B$  is the bounding box of the area  $A$ .

Table IV

mRMR rank of the 16 most discriminative features for the watershed, the GVF and the ACM segmentation<sup>(4)</sup>

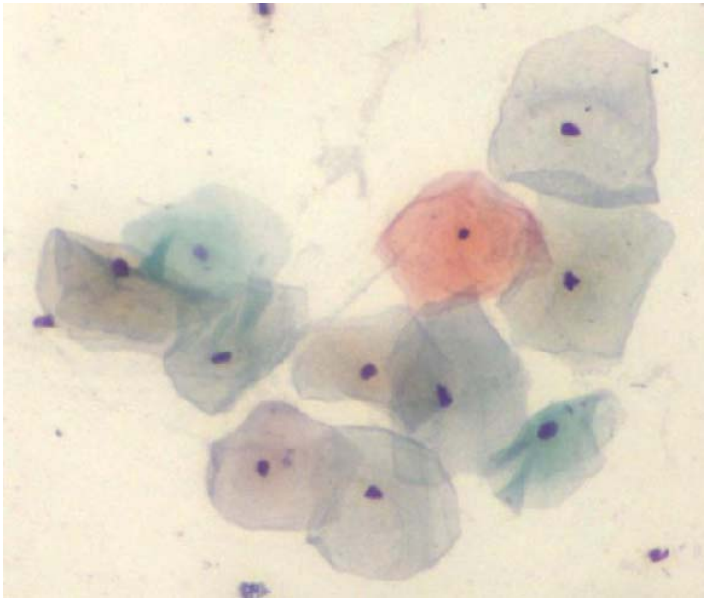
Watersheds		GVF	ACM
1.	Entropy of $B$ in green	<b>Foreground-Background contrast in green</b>	<b>Foreground-Background contrast in red</b>
2.	Perimeter	Minor Axis Length	Minor Axis Length
3.	<b>Foreground-Background contrast in red</b>	Third moment of $A$ in blue	Uniformity of $Ngh$ in green
4.	<b>Std Histogram <math>LBP_{hyperbola}^{riu2}</math> in green</b>	<b>Std Histogram <math>LBP_{hyperbola}^{riu2}</math> in red</b>	<b>Std Histogram <math>LBP_{hyperbola}^{riu2}</math> in red</b>
5.	<b>Circularity</b>	Entropy of $Ngh$ in red	Smoothness of $Ngh$ in green
6.	<b>Foreground-Background contrast in green</b>	Mean Histogram $LBP_{circle}^{riu2}$ in green	Eccentricity
7.	Mean Histogram $LBP_{circle}^{riu2}$ in blue	<b>Foreground-Background contrast in blue</b>	<b>Foreground-Background contrast in green</b>
8.	Entropy of $B$ in red	Eccentricity	Mean Histogram $LBP_{circle}^{riu2}$ in blue
9.	<b>Mean Histogram <math>LBP_{hyperbola}^{riu2}</math> in blue</b>	<b>Mean Histogram <math>LBP_{hyperbola}^{riu2}</math> in blue</b>	<b>Mean Histogram <math>LBP_{hyperbola}^{riu2}</math> in blue</b>
10.	Smoothness of $B$ in green	Uniformity of $B$ in green	Third moment of $A$ in red
11.	Std Histogram $LBP_{circle}^{riu2}$ in red	<b>Foreground-Background contrast in red</b>	<b>Circularity</b>
12.	Entropy of $A$ in green	Std Histogram $LBP_{circle}^{riu2}$ in red	<b>Foreground-Background contrast in blue</b>
13.	<b>Foreground-Background contrast in blue</b>	<b>Std Histogram <math>LBP_{hyperbola}^{riu2}</math> in green</b>	<b>Std Histogram <math>LBP_{hyperbola}^{riu2}</math> in green</b>
14.	<b>Std Histogram <math>LBP_{hyperbola}^{riu2}</math> in red</b>	<b>Circularity</b>	Entropy of $Ngh$ in red
15.	Smoothness of $A$ in red	Entropy of $B$ in green	Mean Histogram $LBP_{circle}^{riu2}$ in red
16.	Third moment of $Ngh$ in blue	Third moment of $Ngh$ in red	Third moment of $A$ in blue

<sup>(4)</sup>  $A$  is the enclosed detected boundary area,  $B$  is the bounding box of  $A$  calculated as the maximum rectangle that contains the detected region,  $B$  and the neighborhood  $Ngh$  is defined as  $Ngh = A^c \cap B$  (see Fig. 5 and text for a detailed description). The features highlighted in bold face fonts are common for the three segmentation techniques and they appear in the first 14 positions in each mRMR rank.

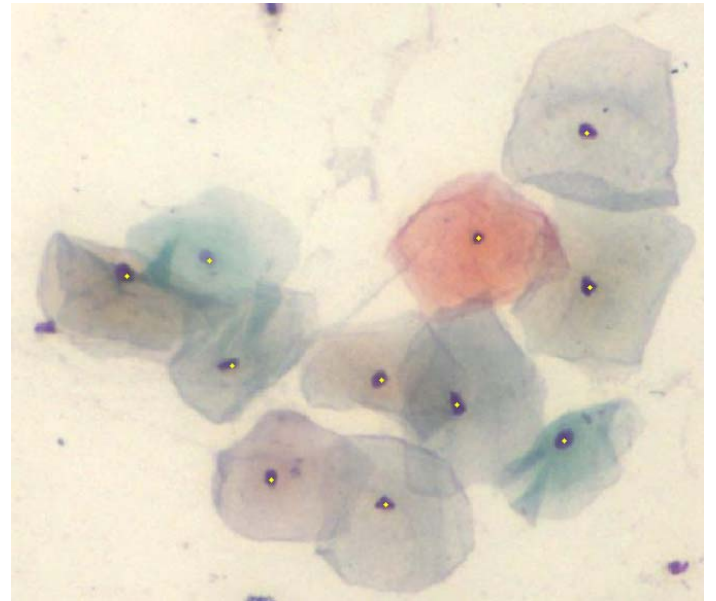
TABLE V

## Clustering Performance

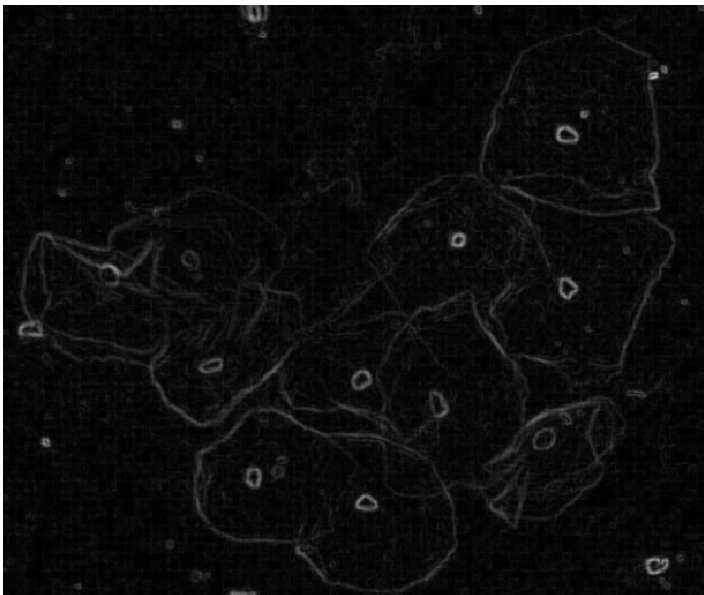
	K-means		Spectral Clustering		SVM	
	global mRMR	leave-one-out mRMR	global mRMR	leave-one-out mRMR	global mRMR	leave-one-out mRMR
Watersheds	84.09%	84.36%	82.64%	82.93%	82.46%	82.52%
ACM	80.09%	79.64%	76.84%	77.00%	81.87%	81.95%
GVF	77.83%	78.76%	77.20%	77.33%	80.20%	80.28%



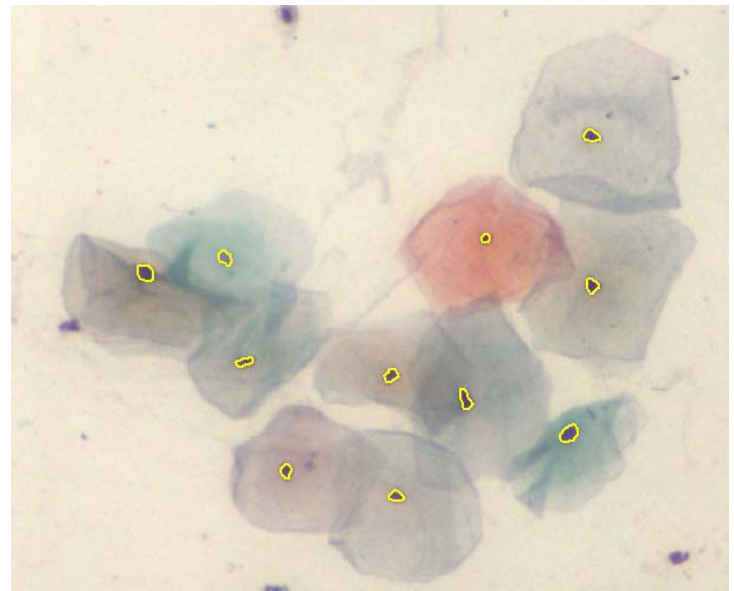
(a)



(b)



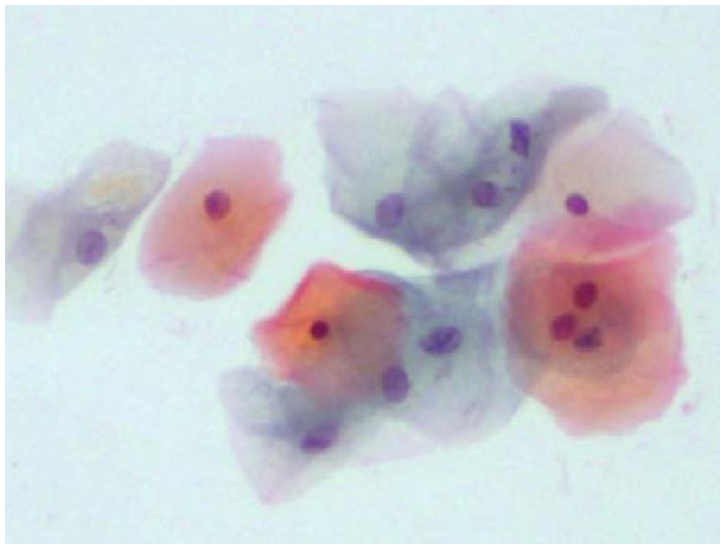
(c)



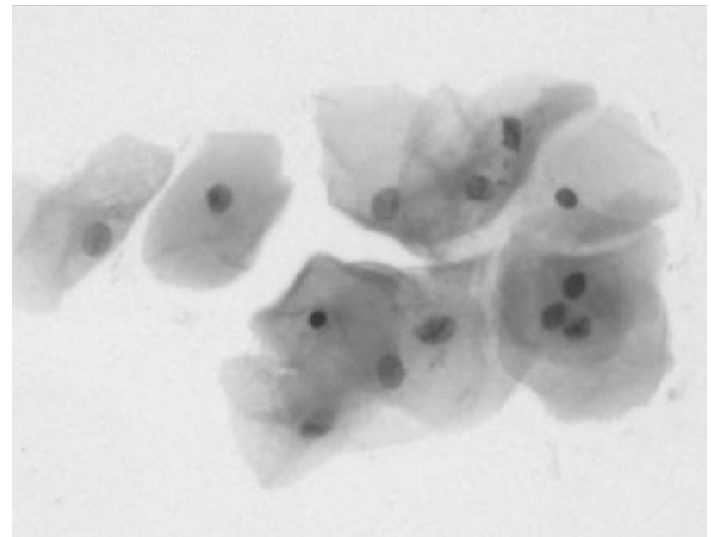
(d)

Fig. 1: (a) Initial image of overlapped cells, (b) the detected nuclei markers, (c) the corresponding color morphological gradient image, (d) the watershed segmentation.





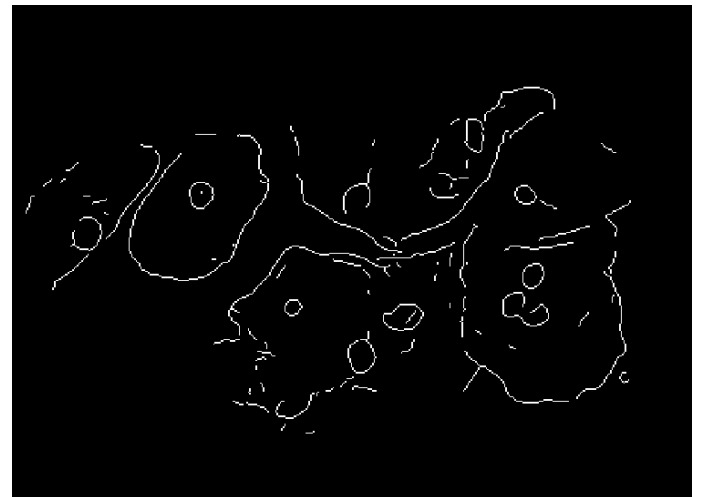
(a)



(b)



(c)



(d)

Fig. 2: (a) Initial image of overlapped cells and (b) the corresponding grayscale image, in which we apply the Canny edge detector. Using a small threshold results in (c) an image with many undesired edges, while using a high threshold results in (d) an image with several significant edges missing.

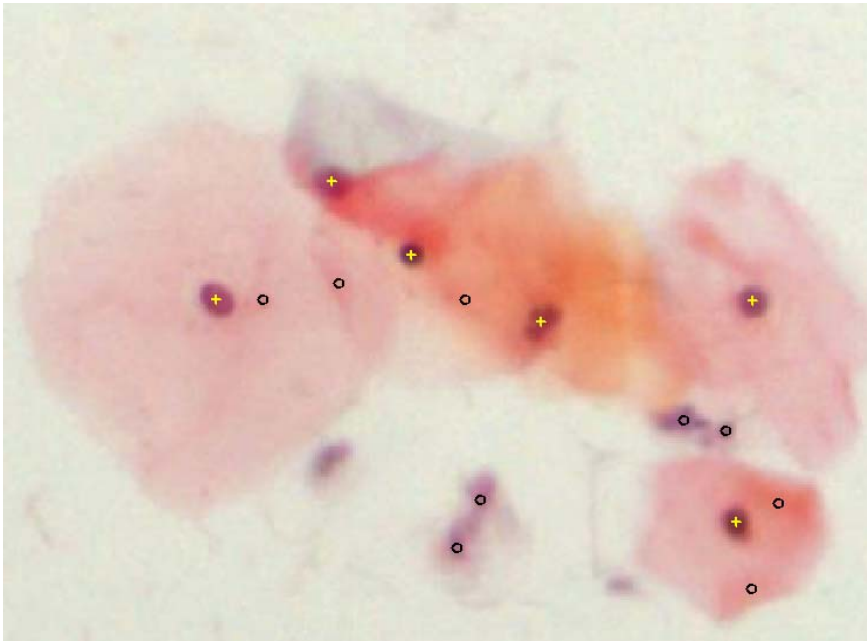
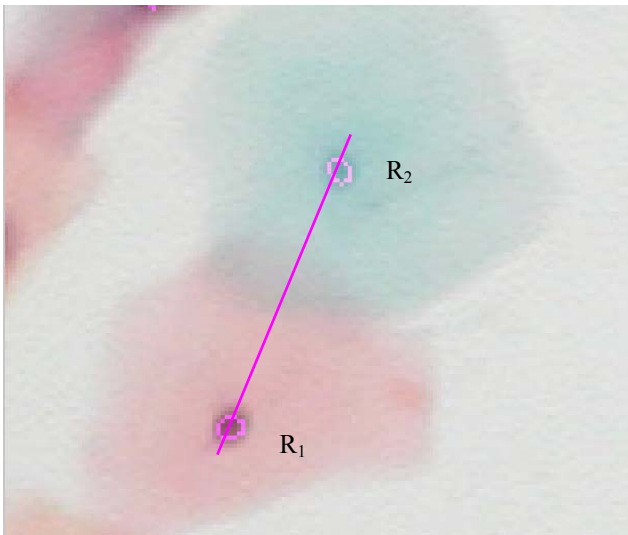
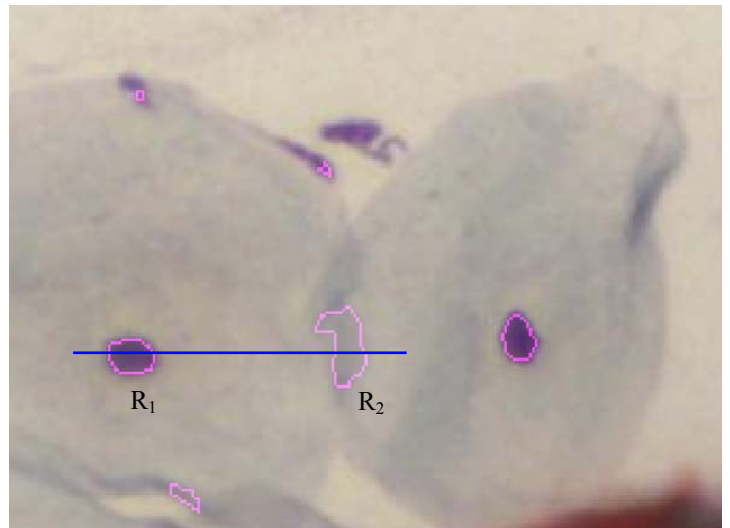


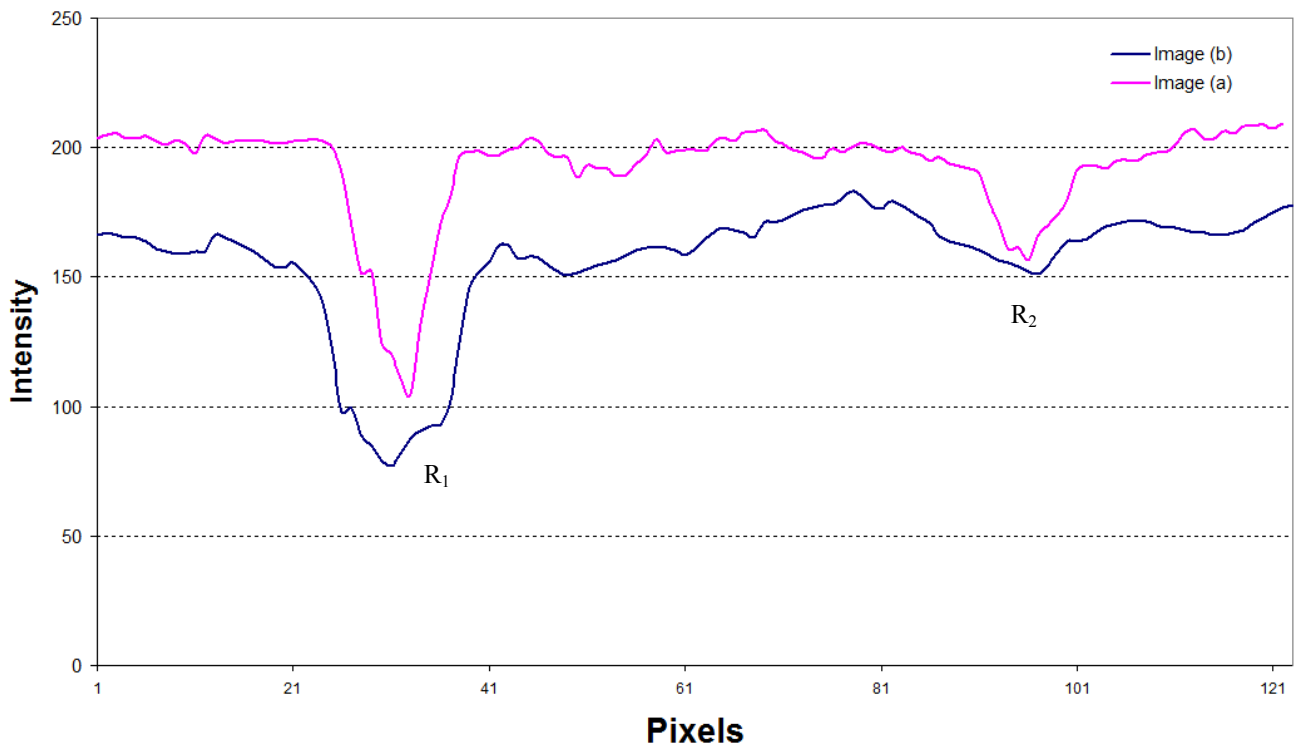
Fig. 3: The detected centroids of the regional minima in the image. The true nuclei locations are represented by a yellow cross and the false positive findings are represented by a black circle.



(a)

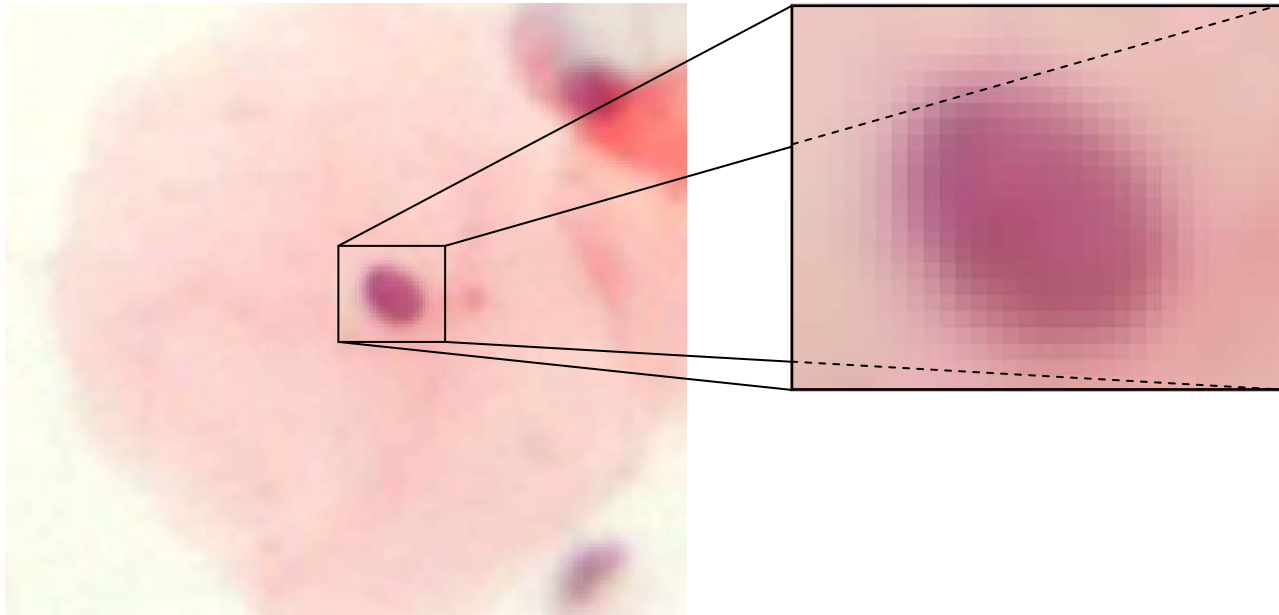


(b)

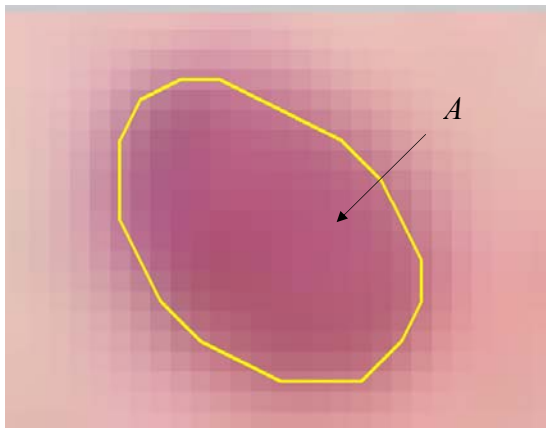


(c)

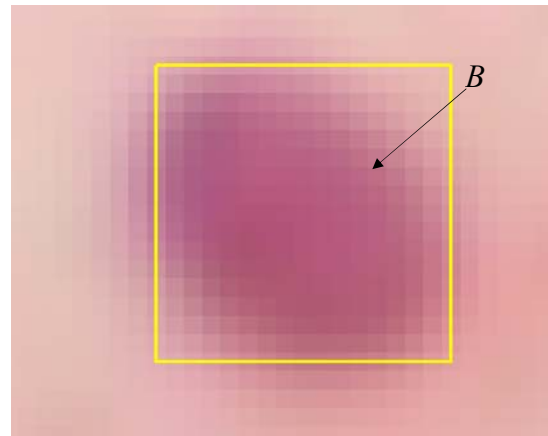
Fig. 4: (a)-(b) The result of the watershed transform in parts of two different cell images. The regions  $R_1$  and  $R_2$  that are detected in both images with the watershed transform are joined with a line for better visualization purposes. In (a) the detected areas  $R_1$  and  $R_2$  correspond to the areas of true nuclei, while in (b) the detected area  $R_1$  corresponds to a nucleus and the area  $R_2$  corresponds to a cytoplasm overlapping area. The variation of the average color image intensity value along the line which joins the areas  $R_1$  and  $R_2$  is depicted in (c). Notice that for the area  $R_1$  we observe sharp reduction of the intensity value in both images. For the area  $R_2$ , although the average intensity value is similar in both images, sharper intensity reduction (in relation with its neighborhood pixels) occurs only for the true nucleus in image (a). This indicates that the use of the neighborhood of each detected area contributes in the recognition of the true nuclei.



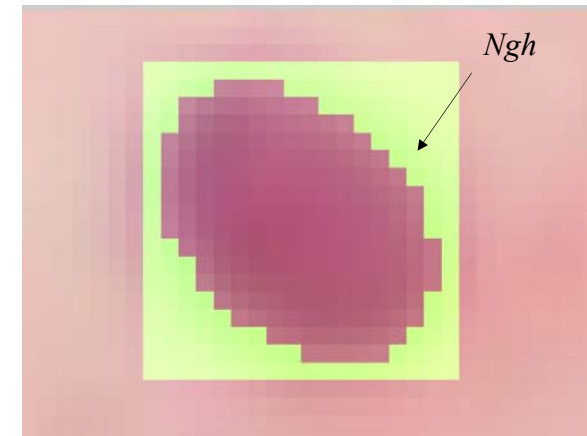
(a)



(b)



(c)



(d)

Fig. 5. The selected areas for the construction of the feature set. (a) A cell from the initial image, (b) the detected nucleus boundary with the watershed transform and the enclosed area A, (c) the area B of the bounding box of the detected boundary, (d) the area of the neighborhood  $Ngh$  ( $A^c \cap B$ ) of the detected nucleus.

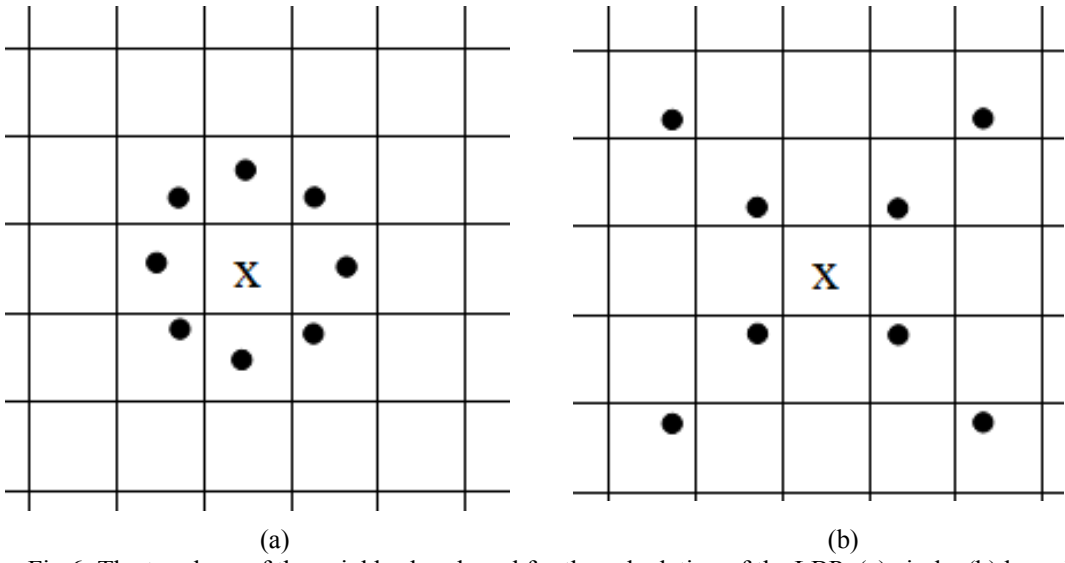


Fig.6: The topology of the neighborhood used for the calculation of the LBP: (a) circle, (b) hyperbola.

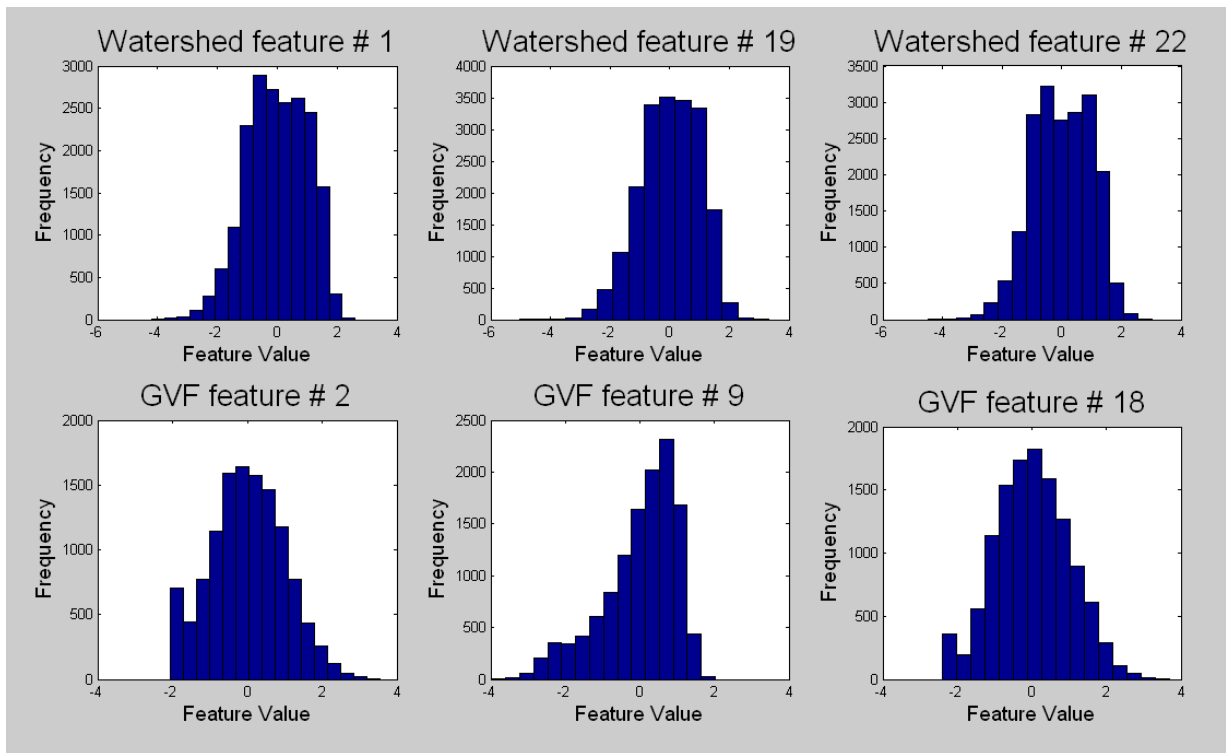
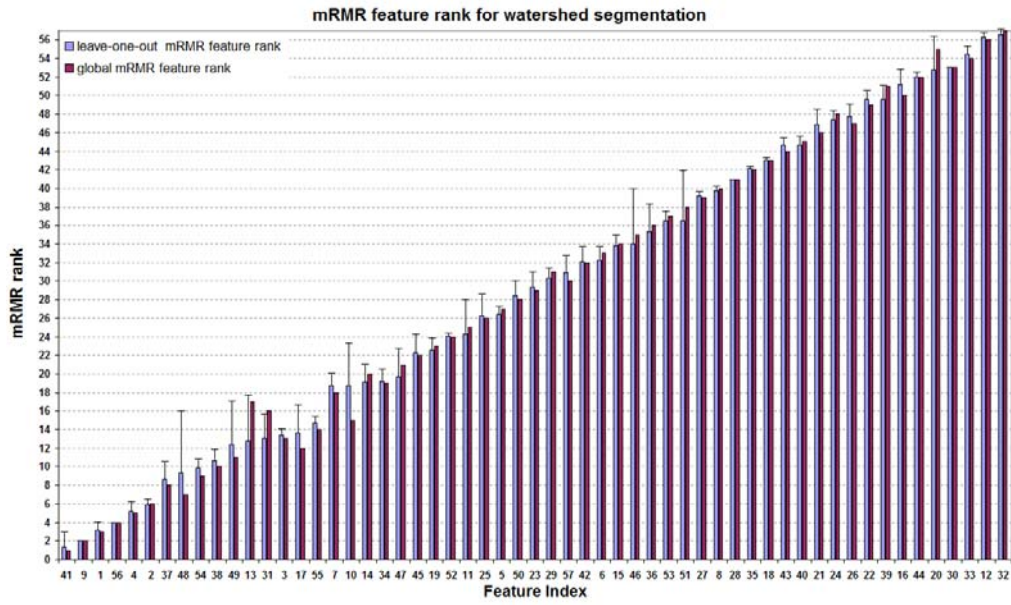
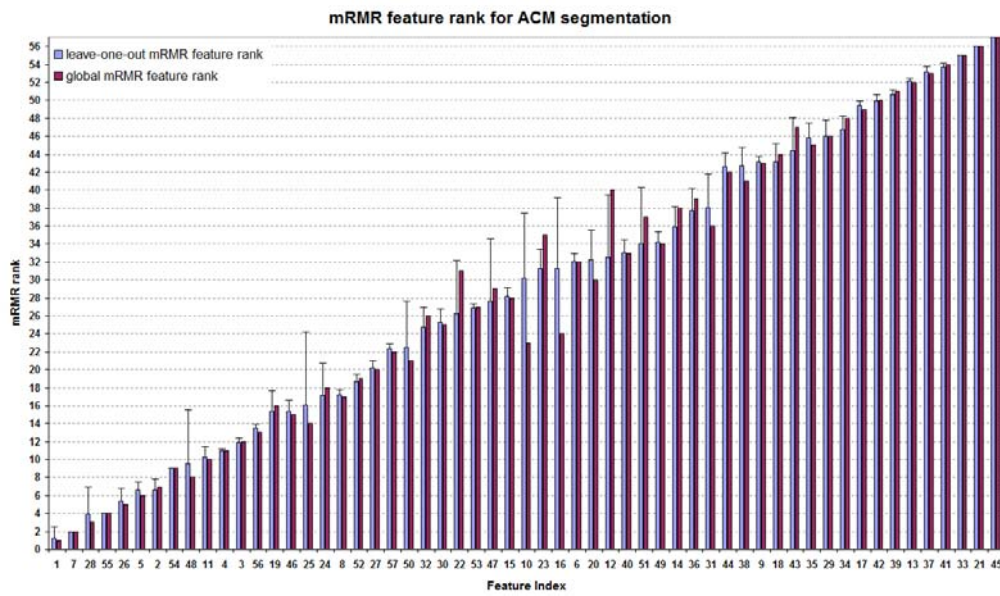


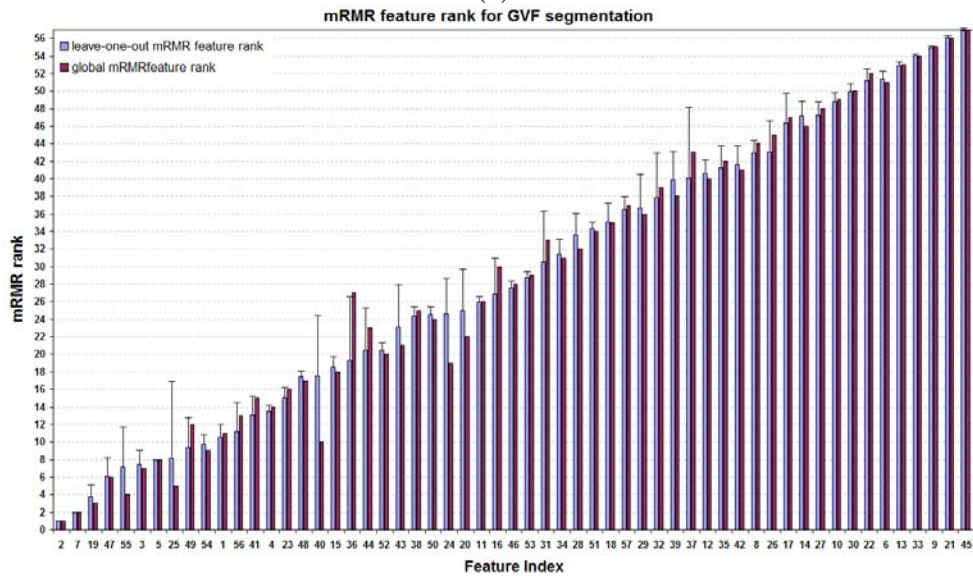
Fig. 7: Representative histograms of some features of the watershed and the GVF segmentation. Notice that their distribution consists of a single blob and this allows their discretization into three states at the positions  $\mu \pm \sigma$ .



(a)



(b)

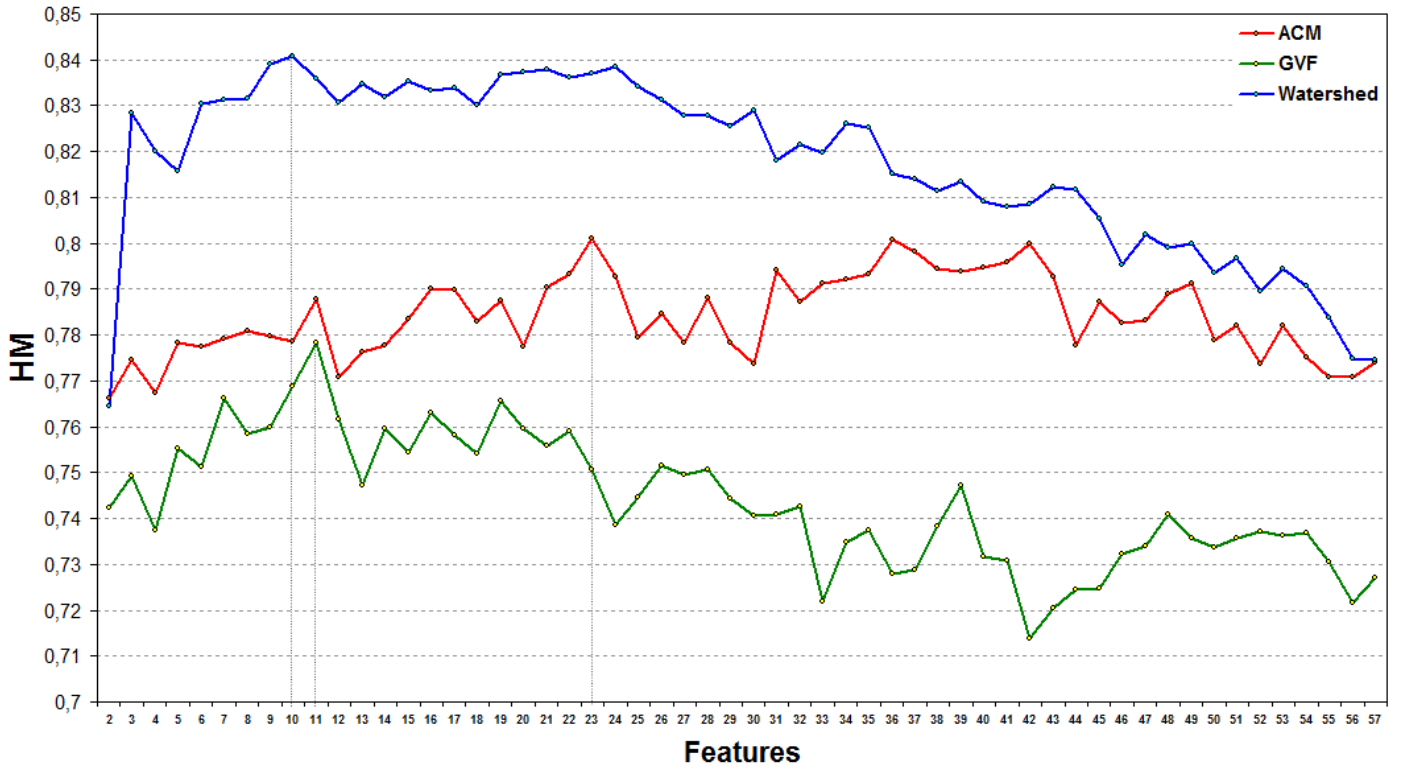


(c)

Fig. 8: The leave-one-out and global mRMR feature rank for the watershed, SCM and GVF segmentation algorithms. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars.

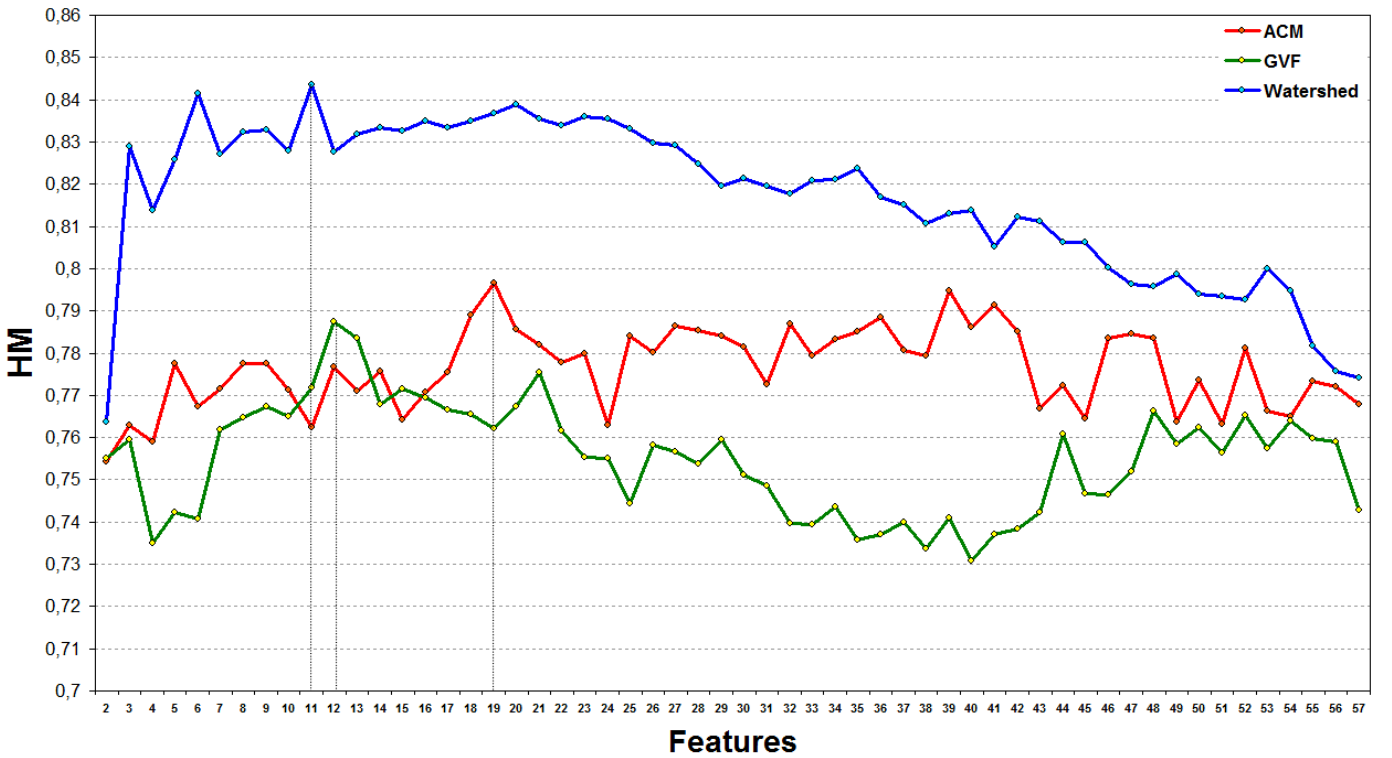


### K-means Clustering (global mRMR rank)



(a)

### K-means Clustering (Leave-one-out mRMR rank)

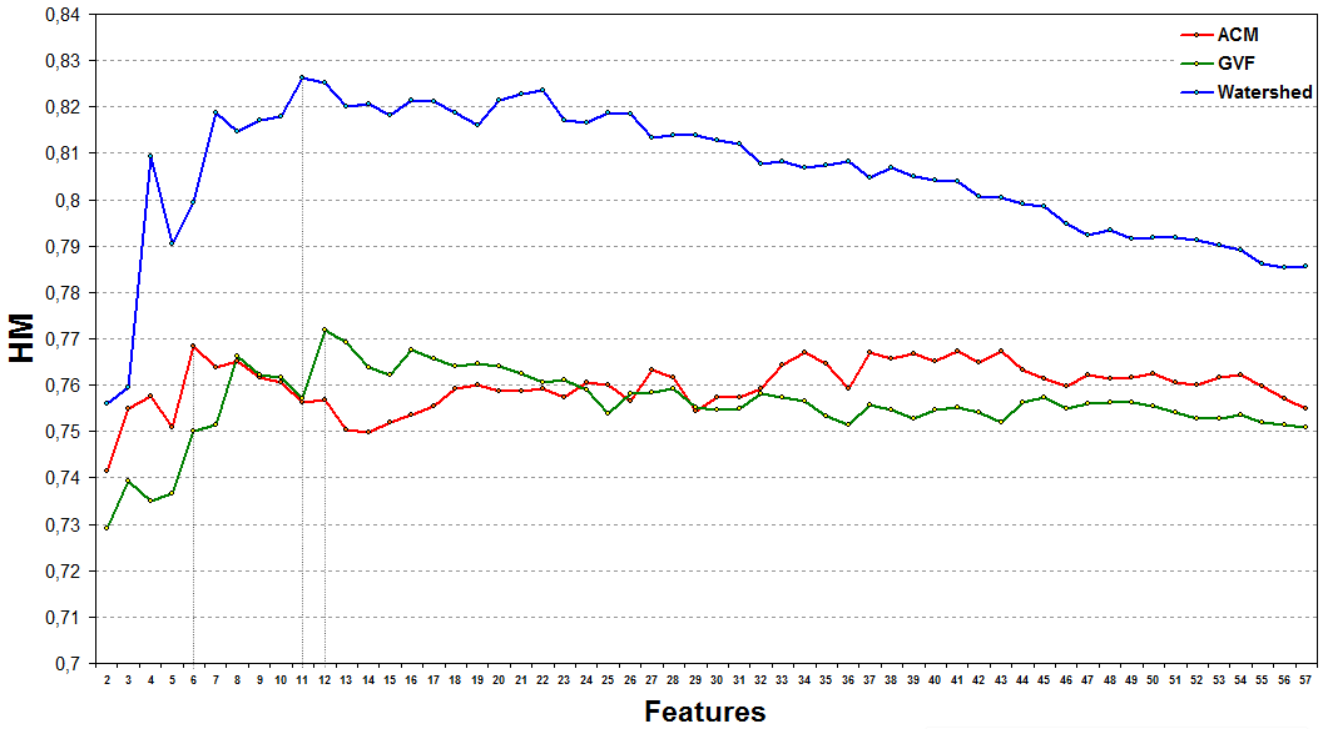


(b)

Fig. 9: Results in terms of the HM measure for the K-means clustering for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. The vertical line indicates the number of features where the HM measure takes its maximum value for the three segmentation methods. These values of HM are contained in Table V.

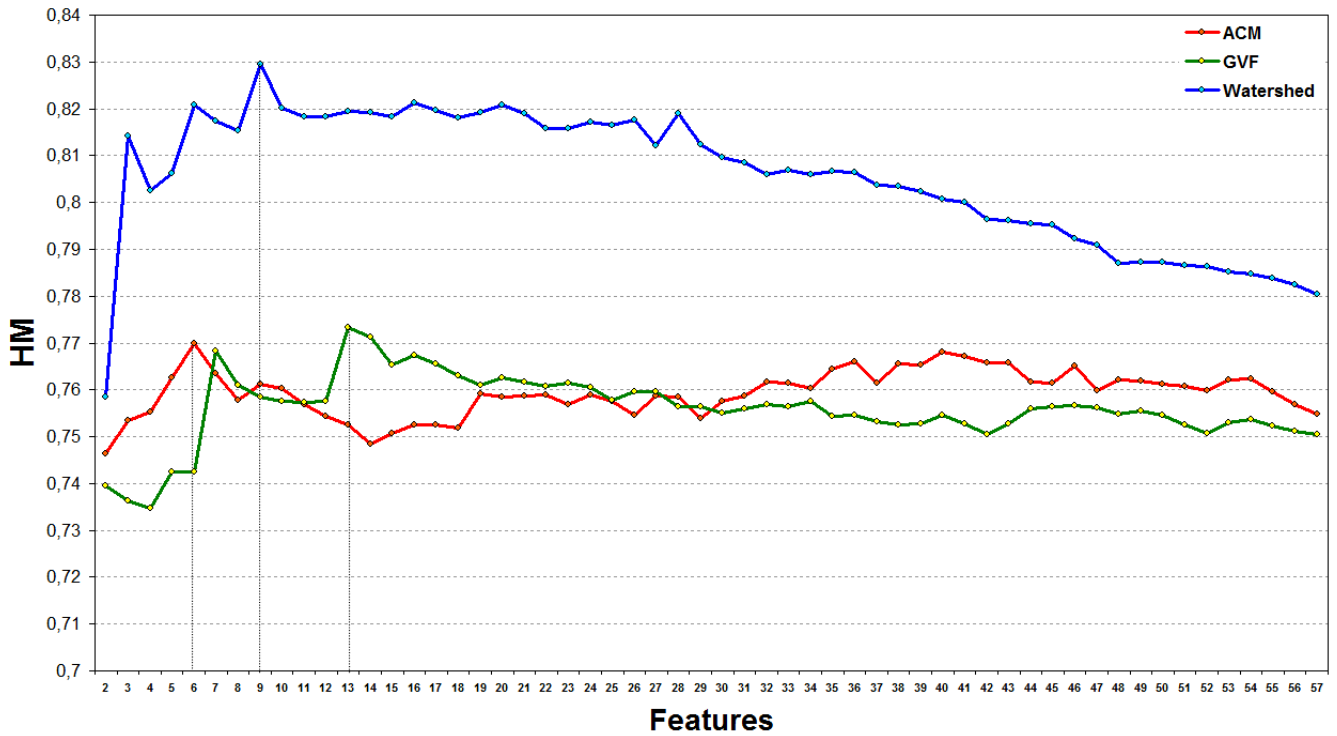


**Spectral Clustering (global mRMR rank)**



(a)

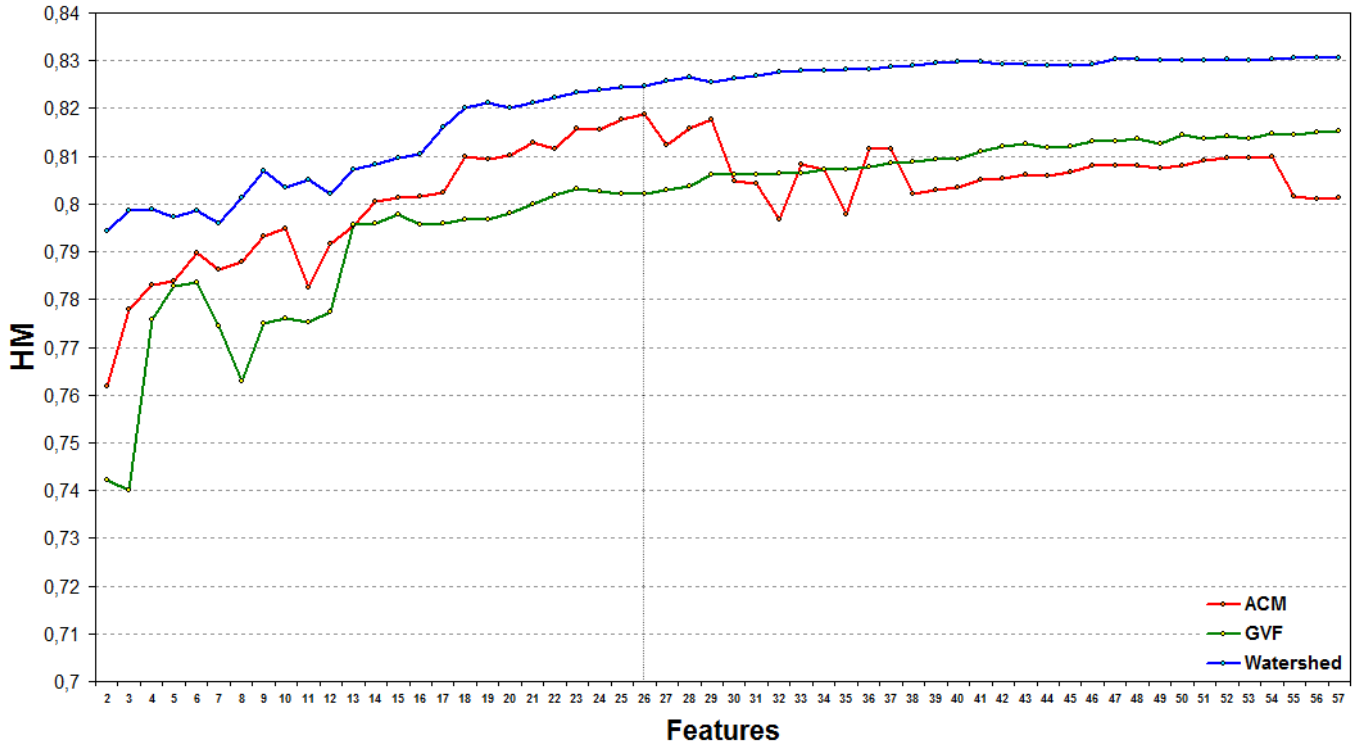
**Spectral Clustering (Leave-one-out mRMR rank)**



(b)

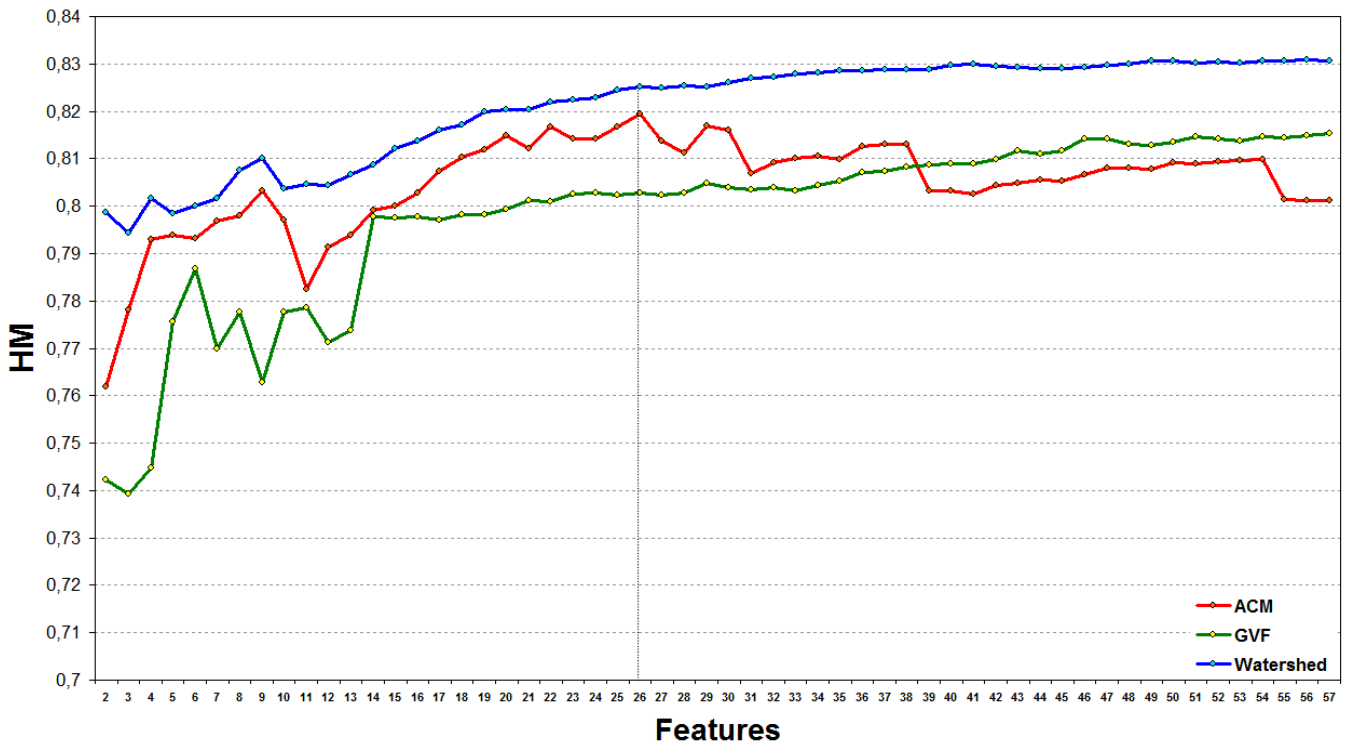
Fig. 10: Results in terms of the HM measure for spectral clustering for ACM, GVF and watershed segmentation for both global (a) and leave-one-out (b) mRMR rank. The vertical line indicates the number of features where the HM measure takes its maximum value for the three segmentation methods. These values of HM are contained in Table V.

### SVM Clustering (global mRMR rank)



(a)

### SVM Clustering (Leave-one-out mRMR rank)



(b)

Fig. 11: Results in terms of the HM measure for the SVM clustering for ACM, GVF and watershed segmentation for both global (a) and leave-one-out (b) mRMR rank. For comparison purposes, the indicative values for HM measure were evaluated using the first 16 features. These features are described in Table IV, while the values of HM are contained in Table V.

	GVF	ACM	Watersheds	Ground Truth
(a)				
(b)				
(c)				

Fig. 12: (a)-(c) Segmentation results for several detected nuclei.


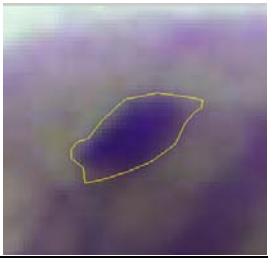
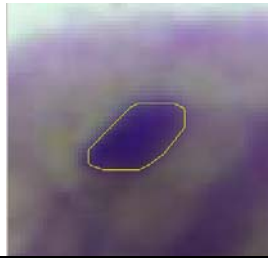
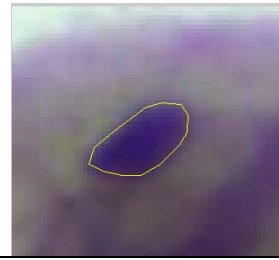




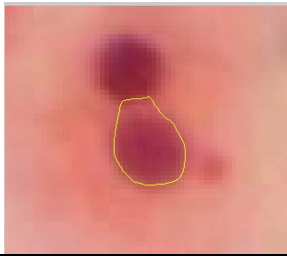
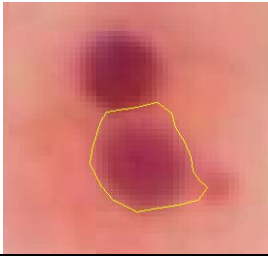
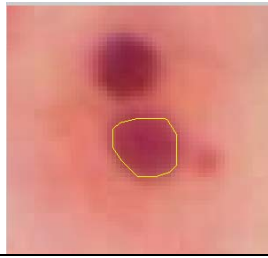
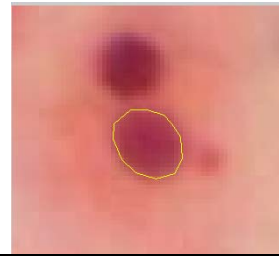
	GVF	ACM	Watersheds	Ground Truth
(a)				
(b)				
(c)				

Fig. 13: Representative cases of failure for ACM and GVF segmentation in images with (a) weak gradient at the nucleus boundary, (b) the inhomogeneities of the nucleus intensity and (c) the existence of high value of gradient in the neighborhood of the nucleus boundary .