Automatic detection of cervical cells in Pap-smear images using polar transform and k-means segmentation

Mihai Neghina[1], Christoph Rasche[2], Mihai Ciuc[2], Alina Sultana[2] and Ciprian Tiganesteanu[3]
[1]University Lucian Blaga of Sibiu
e-mail: mihai.neghina@gmail.com
[2]University Politehnica of Bucharest
e-mail: rasche15@gmail.com, mihai.ciuc@upb.ro, asultana@imag.pub.ro
[3]Genetic Lab SRL
e-mail: ciprian.tiganesteanu@yahoo.ro

***Abstract* — We introduce a novel method of cell detection and segmentation based on a polar transformation. The method assumes that the seed point of each candidate is placed inside the nucleus. The polar representation, built around the seed, is segmented using k-means clustering into one candidate-nucleus cluster, one candidate-cytoplasm cluster and up to three miscellaneous clusters, representing background or surrounding objects that are not part of the candidate cell. For assessing the natural number of clusters, the silhouette method is used. In the segmented polar representation, a number of parameters can be conveniently observed and evaluated as fuzzy memberships to the non-cell class, out of which the final decision can be determined. We tested this method on the notoriously difficult Pap-smear images and report results for a database of approximately 20000 patches.**

***Keywords* — Pap-smear images, Cell detection, Polar representation, k-means clustering, Fuzzy membership.**

I. Introduction

One of the most widely used techniques for preparing medical samples in order to identify cervical cancer employs a staining technique developed by Papanicolaou [1], resulting in the so-called Pap smear images. The detection of cells from such cytological images, the first steps in the analysis of samples, is therefore a difficult but important problem for the medical community.

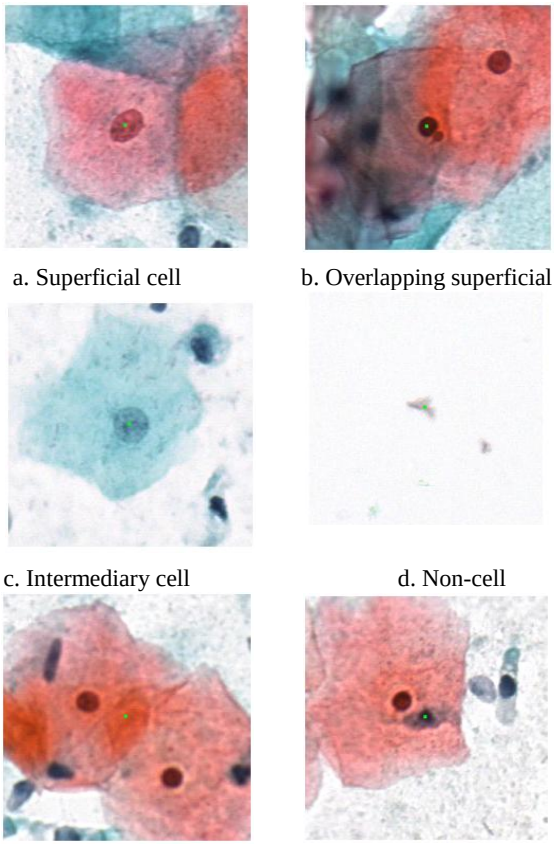
Although there have been various attempts to automatically extract cells from Pap smear images, typically using methods involving thresholds [2][3], morphological operations [4][5] or gradient vector flow snakes [6][7], we think a different approach also merits consideration, namely one based on polar transforms around nuclei and k-means for segmentation. Polar representations have already been used in cell analysis. Angulo [8] used a log-polar transform in conjunction with morphological operations (skeleton) for the analysis of the shape of erythrocytes, while Nosrati et al. [9] employed the polar transform for the computation of directional derivatives in an attempt to separate overlapping cervical cells. Our study proposes the use of the polar transform not only as a convenient representation, but also as the basis for a dimensionality reduction, for unsupervised clustering and for feature extraction. In this study, seed points are selected by Rasche's et al. [10] method using geometric analysis of iso- and edge contours, but other seed points could possibly serve equally well. The segmentation provides for each candidate an estimate of the quasi-nucleus, the quasi-cytoplasm and other clusters (e.g. background sections, etc.) – quasi being used here to stress that not all candidates are cells therefore not all candidate images contain a nucleus or a cytoplasm. Several cell candidate parameters are then computed and their fuzzy membership to the non-cell class is used to decide if the candidate is not a cell.

The remainder of the paper is divided as follows: section II describes the database, section III describes the method of segmentation of candidate images, section IV describes the decision module and section V presents the results and perspectives.

II. Database

Medical sample images are obtained with a scanner by VENTANA iScan Coreo, at an optic zoom of 40, generating images of approximately 75000 x 75000 pixels; at that resolution, a nucleus has a diameter between 30 and 60 pixels (depending on cell type) and a cell has a maximal diameter of 300 pixels.

Using the iso-contours method on smear images from five subjects, a number of 20148 candidate images have been extracted: 9405 actual cells (8399 of which are centred on the nucleus and 1006 are centred outside the nucleus) and 10743 non-cells. Figure 1 shows six examples of candidate images (301x301 pixels). The green dot in the centre of each image, representing the seed for the proposed method, has been added artificially in order to enable easy human assessment.



a. Superficial cell b. Overlapping superficial cells

c. Intermediary cell d. Non-cell

e. Cell seeded on cytoplasm f. Cell seeded on artefact

Fig. 1. Cell candidate images with marked seeds (green dot).

III. Segmentation of candidate images

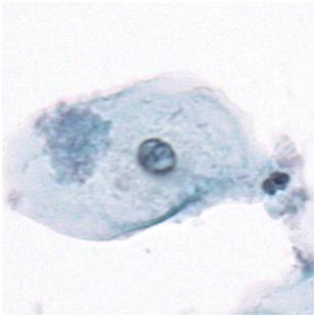
Seed points are detected by the method described in Rasche et. al [10], by extracting iso-contours at different intensity levels - a method different from and faster than active contours. Most nuclei display an iso-contour but there are many other iso-contours in an image. In order to select iso-contours that represent potential nuclei candidates, a number of criteria are applied, such as minimum and maximum diameter, minimum contrast, a minimum degree of roundness etc.

The automatic segmentation method proposed in this study is composed of three steps:

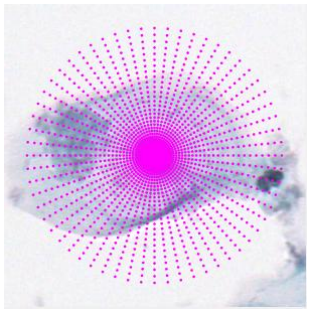
- *A.* An image Cartesian-to-Polar transformation, re-dimensioning and the addition of a fourth pixel dimension;
- *B.* k-means as the primary (unrefined) segmentation
- *C.* Post-processing, including merging of clusters

A. Image transformation

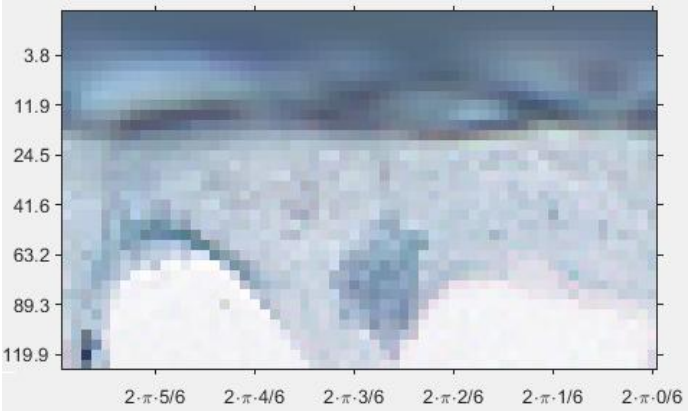
The candidate images contain, in the ideal case, concentric layers (representing nucleus, cytoplasm and background) and therefore the polar transformation comes as a natural first step. The transformation itself does not introduce or uncover new information, but it re-arranges the information in a way that makes further operations easier. Furthermore only pixels at certain positions (certain angles theta and certain distances d from the seed) are kept in the transformed image. These positions marked with magenta in Fig. 2. For most positions, the magenta points fall in-between pixels and therefore the nearest-neighbour pixel values are picked. Although the bilinear interpolation gives a smoother appearance, it offers relatively little advantage over the nearest-neighbour approach. Compared to the bilinear interpolation, the nearest-neighbour approach results in more misclassified isolated pixels. However, these pixels are labelled *black* (unused) in step III.C and ignoring them does not change the indicators in chapter IV in any significant way.



a. Candidate image



b. Sampling points



c. Polar transform of the re-sampled candidate image

Fig. 2. Example of a Cartesian-to-Polar transform with marked sampling points (magenta) on the original image

Using a set of 60 equally spaced angles and 36 distances, the number of pixels can be reduced from 90601 to just 2160 while at the same retaining the most significant elements of the image. The number of angles and distances are adjustable parameters of the proposed method, but the chosen values seemed to be sufficient for the current resolution. Since the most significant information is concentrated around the nucleus (and the nucleus is assumed to contain the seed), the selected distances become more sparse with distance from seed.

In order to improve the k-means segmentation in step B, a fourth component is added pre-emptively to each pixel of the transformed image. The reason for the fourth value is to convey the information to the k-means clustering that, ideally, the top pixels form the nucleus, the middle pixels form the cytoplasm and the lower pixels form the background of the cell. Looking at the values v along any column of the transform image, the fourth dimension encourages pixels at very small or very large distances from the seed to classify naturally, while also gently increasing the Euclidian distance between the top and bottom pixels in the transformed image. The fourth dimension values are computed as values of a scaled logarithmic sigmoid, where k_s is the scaling factor and $f(d)$ is a linear function of distance to the seed:

$$v = k_s / \left(1 + e^{-f(d)}\right) \tag{1}$$

The function f is given generic so that the values of the exponent can be adjusted according to the microscopic resolution of the scanner. In our case, the maximal radius of the cell is 150 pixels and the minimal radius of the nucleus is 30 pixels. Because the values of the *RGB* layers are scaled to the interval [0...1] for all further processing, the values of v obey the same constraints:

$$f(0) = -3.5 \qquad v(0) \approx 0 * k_s \tag{1.1}$$

$$f(30) = -0.7 \qquad v(30) \approx 0.1 * k_s \tag{1.2}$$

$$f(150) = 3.5 \qquad v(150) \approx 1 * k_s \tag{1.3}$$

Small scaling factors k_s would have an unnoticeable effect on the segmentation because top and bottom pixels would have roughly the same 4th value, whereas large scaling factors k_s may overshadow the other three dimensions (the actual colour components of the pixels) and force an unnaturally ordered segmentation. Figure 3 shows the results of *k*-means clustering and the scaling factor chosen, as well as the effects of extreme (very small and very large) scaling factors.

a. Polar transform

b. Clustering with $k_s = 0$

c. Clustering with $k_s = 0.35$

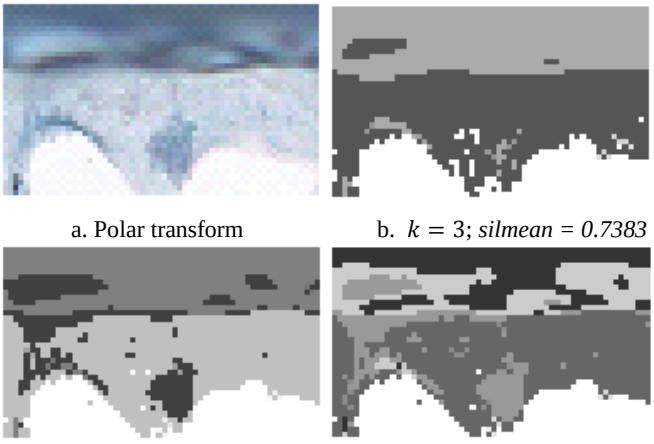
d. Clustering with $k_s = 1$

Fig. 3. Example of a candidate image and the results of *k*-means clustering for various scaling factors k_s

The appropriate scaling factor k_s is computed by averaging the standard deviation of the uppermost 50% of the rows in the transformed image, since uneven rows with larger deviations require more forcing values on the 4th dimension. Factor k_s is capped at 0.6 to avoid the overshadowing of the colour dimensions.

B. k-means clustering

The 4D pixel array is clustered using the *k*-means for three values of *k* (*k=3*, *k=4* and *k=5*) because we expect at least three clusters, but we tolerate up to 2 more clusters of 'noise' regions. More than 5 clusters leads to over-segmentation. For determining the optimal number of clusters, we have considered the silhouette method [11], denoted *silmean* in Fig. 4.



a. Polar transform b. $k = 3$; *silmean = 0.7383*

c. $k = 4$; *silmean = 0.7228* d. $k = 5$; *silmean = 0.6745*

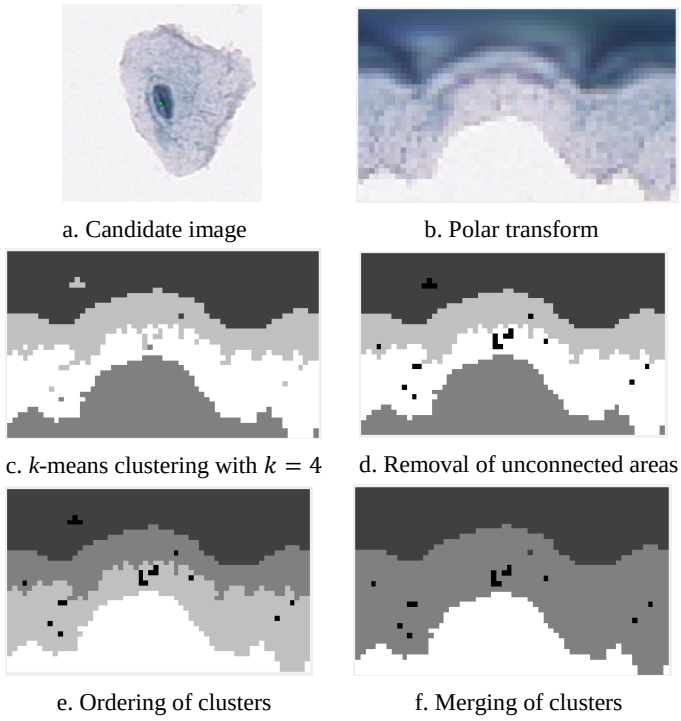
Fig. 4. *k*-means clustering results for various number of classes *k*

C. Post-processing

This step consists of various morphological operations that further improve the clustering results.

- *Removal of unconnected areas.* Considering that the cell should have the nucleus and cytoplasm in one connected area each, this step keeps for each cluster only the largest connected area and re-assigns any other group of pixels to label *black* (unused), represented as black in further images. For this step, the first and last columns of the transformed image are considered connected (since they represent consecutive sampling angles).
- *Merging of nucleus parts.* If the centroids of the uppermost two clusters are close enough, they should merge. Also as part of this step, black areas engulfed by the nucleus are assigned the same label, the nucleus label.
- *Ordering of clusters*. This step orders the clusters so that the uppermost cluster (usually the nucleus) is the darkest and the lowermost cluster (usually the whitish background) is the lightest. As decided at the previous step, the nucleus is the cluster with the lowest mean vertical position of the pixels (closest to the uppermost row). Cytoplasm is the cluster with the greatest contact to the nucleus. The other clusters are then sorted by the mean vertical position of the pixels.
- *Merging of cytoplasm parts.* For the cytoplasm cluster determined at the previous step and each of the other clusters except the nucleus, the statistical mean and standard deviation of the *red* layer or the *blue* layer are computed (depending on the dominant colour of the nucleus). If the normalized distributions of the cytoplasm cluster and another cluster overlap significantly, the two clusters are merged.

Figure 5 shows the full chain of processing and the results of segmentation operations for a given candidate cell. The final number of clusters after post-processing can be smaller than in the primary segmentation due to cluster merges.



a. Candidate image b. Polar transform

c. *k*-means clustering with $k = 4$ d. Removal of unconnected areas

e. Ordering of clusters f. Merging of clusters

Fig. 5. Processing chain for a candidate image: the computed scaling factor was $k_s = 0.56$ and the best average silhouette value *silmean = 0.7322* was obtained by *k*-means with *k = 4* clusters.

IV. Cell detection module

Once the segmentation has been completed, parameters useful to the decision making module can be extracted. The aim of the decision module is thus the identification of non-cells with the reasoning that anything that is not weird enough should be treated as an actual cell. There are two types of parameters: strong indicators and weak indicators that the candidate is a non-cell. For each indicator, we compute a membership value to the class non-cell $\boldsymbol{\mu_i}$. The final decision, described in section IV.C, is taken by comparing the aggregate of those membership values to a pre-defined threshold $\boldsymbol{T}$.

A. Extraction of strong indicators

The strong indicators are:

I_1: Whiteness of the nucleus class
I_2: Whiteness of the cytoplasm class
I_3: Contrast between the nucleus class and the cytoplasm class

These are considered strong indicators because any of them can single-handedly force the decision that the candidate is a non-cell, as described in the decision rule (section IV.C).

For the strong indicators, the whiteness measure for both the nucleus class and the cytoplasm class is computed as the average between the *red* and *blue* layers of all pixels belonging to the respective class. The green layer is irrelevant when both *red* and *blue* layers have high values: a large value for green would make the pixels whitish and a low value for green would make the pixels magenta, but in both cases it would indicate that the candidate is a non-cell. The fuzzy membership to the non-cell class for the whiteness indicators is represented in Fig. 8.a.

The contrast between the cytoplasm class and the nucleus class is simply the difference between the whiteness measures of the respective classes. The fuzzy membership to the non-cell class for the whiteness indicators is represented in Fig. 8.b.

B. Extraction of weak indicators

The weak indicators that the candidate is a non-cell are:

I_4: A radius estimation for the nucleus cluster
I_5: A measure of nucleus cluster roundness
I_6: A contact index between the nucleus and the cytoplasm cluster
I_7: The area ratio between the nucleus and the cytoplasm cluster
I_8: A nucleus colour index

Ideally, the nucleus is round, but the centre of the candidate image is not necessarily the centre of the nucleus. The Power-of-a-point theorem (Euclid Elements, III.35) offers a convenient way of estimating the lower bound of the nucleus radius, as well as a way to compute a measure of the nucleus roundness.

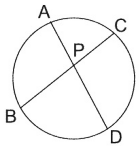


Fig. 6. Power-of-a-point theorem $PA \cdot PD = PB \cdot PC = constant$

For any point P inside a circle, the theorem states that any secant AD that contains point P will be divided into segments PA and PD whose product is constant, as shown in Fig. 6.

However, due to the imperfection of the nucleus roundness, when computing the products of secant segments, the values may not match perfectly. We consider the mean (across all angles) of the square root of these products as our estimate for the radius. This value also is a lower bound for the real radius, in case the nucleus really is a perfect circle. The fuzzy membership to the non-cell class for the radius indicators is represented in Fig. 8.c.

Also, the standard deviation of the square root of these products will be our measure for the roundness of the nucleus. If the nucleus is a perfect circle, the standard deviation should be very close to 0, due to the Power-of-a-point theorem, whereas nuclei with high eccentricity will give a large standard deviation value. These computations regarding the roundness of the nucleus are simplified due to the polar transformation. Since *P* is located at the uppermost row in the transformed image, segments corresponding to *PA* and *PD* are vertical distances from the top of the image to the edge of the nucleus class (separated by π on the x-axis), as shown in Fig. 7.

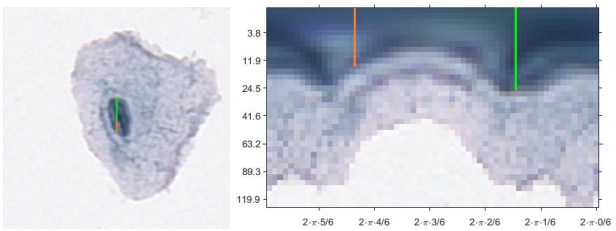


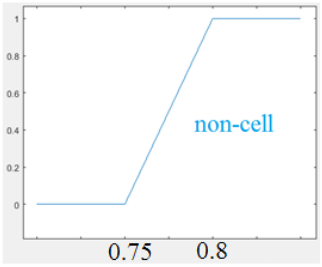
Fig. 7. Representation of a nucleus secant in the polar transform

The fact that the distances are not drawn to scale in the transformed image is taken into account by using distances from the original image (marked on the y-axis). Because there will be a greater variation for larger radii (due to the quantisation process of distances when applying the polar transform), the standard deviation is divided by the square root of the estimated radius (in order to allow a greater tolerance). The membership function to the non-cell class for the nucleus roundness is shown in Fig. 8.d.

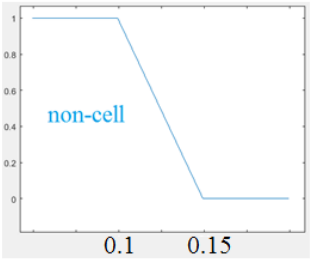
The contact index is computed as the number of angles for which there is contact between the nucleus class and the cytoplasm class, divided by the total number of angles taken into consideration. Once again, the computation of the contact is simplified due to the polar transformation. The membership function to the non-cell class for this parameter is shown in Fig. 8.e.

Before we can compute the area ratio between the nucleus class and the cytoplasm class, we have to compute each area. Due to the segmentation and quantization, discrete points on the boundary of each class are known, so using Gauss's area formula (also known as the Surveyor's formula or Shoelace formula), the right answer can be extracted even from the distorted transformed image. The membership function to the non-cell class for this parameter is shown in Fig.8.f

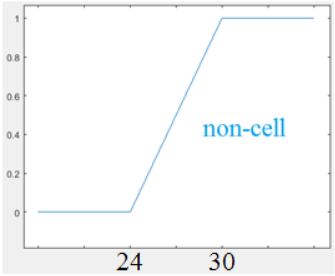
Finally, a nucleus colour index is giving information whether the supposed cell is tainted red, blue or undecided. If the mean values of the red and blue layers in the nucleus class are too close, then either the candidate is not a cell or is a superposition or conglomerate of cells. In both cases, the candidate is regarded as an artefact which cannot be properly analysed as a single cell. The membership function to the non-cell class for this parameter is shown in Fig. 8.g



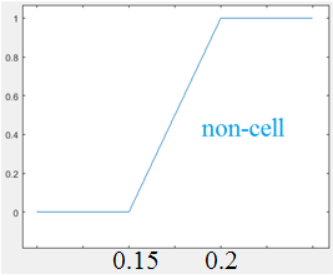
a. Fuzzy membership to class non-cell for the indicators I_1 & I_2 (nucleus & cytoplasm whiteness)



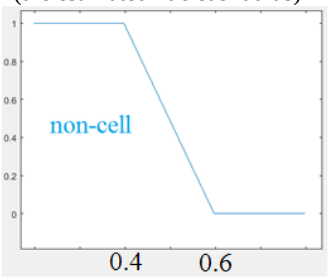
b. Fuzzy membership to class non-cell for indicator I_3 (the nucleus–cytoplasm contrast)



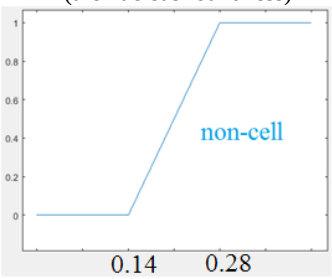
c. Fuzzy membership to class non-cell for indicator I_4 (the estimated nucleus radius)



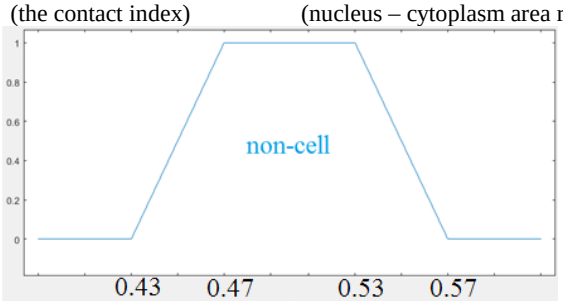
d. Fuzzy membership to class non-cell for indicator I_5 (the nucleus roundness)



e. Fuzzy membership to class non-cell for indicator I_6 (the contact index)



f. Fuzzy membership to class non-cell for indicator I_7 (nucleus – cytoplasm area ratio)



g. Fuzzy membership to class non-cell for indicator I_8 (the nucleus color index)

Fig. 8. Fuzzy memberships to class non-cell for all indicators

C. Final decision for the cell detection module

For each candidate, we compute the membership values to class non-cell $\boldsymbol{\mu_i}$ according to the strong indicators I_1, I_2 and I_3 and the membership values to class non-cell $\boldsymbol{\mu_j}$ according to the weak indicators $I_4 - I_8$.

The aggregate membership $\boldsymbol{\mu}$ is the weighted sum of the fuzzy memberships for all indicators. Strong indicators are weighted with $W_s = 1$, while weak indicators are weighted with $W_w = 0.8$

$$\boldsymbol{\mu} = \sum_{\substack{strong\\indicators\\i}} \boldsymbol{W_s \cdot \mu_i} + \sum_{\substack{weak\\indicators\\j}} \boldsymbol{W_w \cdot \mu_j} \quad (2)$$

A candidate is deemed non-cell if the aggregate membership $\boldsymbol{\mu}$ exceeds threshold $\boldsymbol{T = 1}$:

$$\boldsymbol{\mu} \geq \boldsymbol{T} : \textit{Non-cell} \quad (3)$$

$$\boldsymbol{\mu} < \boldsymbol{T}: \textit{Cell}$$

Because $W_s = 1$, each strong indicator can enforce the decision of non-cell alone, but weak indicator memberships are scaled down and thus need other partial memberships to add up to the threshold $\boldsymbol{T}$. Figure 9 shows several candidates for which the decision was *Cell* or *Non-cell* as well as all corresponding indicator memberships.

V. Results and Perspectives

The segmentation and decision has been applied to ~20000 candidate cells, approximately half of which had been manually labelled as non-cells. Furthermore, 1006 of the candidates labelled as cells have the seed in the cytoplasm (i.e. outside of the nucleus).

In order to evaluate the performances of the method, we have chosen the following measures:

TP = True positives = cells detected as cells

TN = True negatives = non-cells detected as non-cells

FP = False positives = non-cells detected as cells

FN = False negatives = cells detected as non-cells

Correct Detection Rate *[%]*:

$$CDR = \frac{TP}{(TP + FN)} \quad (4)$$

Correct Rejection Rate *[%]*:

$$CRR = \frac{TN}{(TN + FP)} \quad (5)$$

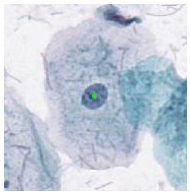
Total Success Rate *[%]*:

$$TSR = (CDR + CRR)/2 \quad (6)$$

Table 1 presents the performances of the proposed method, considering two cases, when the cells that are not seeded in the nucleus are considered as part of class *Cells* and as *Non-cells*.

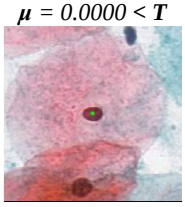
Table 1. Performances of the proposed method

	Considering cytoplasm seeds as cells	Considering cytoplasm seeds as non-cells
TP	6652	6394
TN	8362	9110
FP	2381	2639
FN	2753	2005
CDR[%]	70.73	76.13
CRR[%]	77.84	77.54
TSR[%]	74.28	76.83



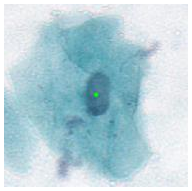
Decision: Cell
$\boldsymbol{\mu}$ *= 0.0000 <* $\boldsymbol{T}$

$\mu_1 = 0.0000$
$\mu_2 = 0.0000$
$\mu_3 = 0.0000$
$\mu_4 = 0.0000$
$\mu_5 = 0.0000$
$\mu_6 = 0.0000$
$\mu_7 = 0.0000$
$\mu_8 = 0.0000$



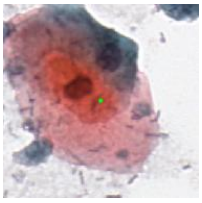
Decision: Cell
$\boldsymbol{\mu}$ *= 0.1399 <* $\boldsymbol{T}$

$\mu_1 = 0.0000$
$\mu_2 = 0.1399$
$\mu_3 = 0.0000$
$\mu_4 = 0.0000$
$\mu_5 = 0.0000$
$\mu_6 = 0.0000$
$\mu_7 = 0.0000$
$\mu_8 = 0.0000$



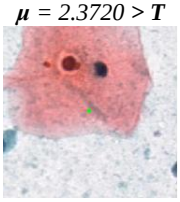
Decision: Cell
$\boldsymbol{\mu}$ *= 0.9239 <* $\boldsymbol{T}$

$\mu_1 = 0.0000$
$\mu_2 = 0.0000$
$\mu_3 = 0.0000$
$\mu_4 = 0.4703$
$\mu_5 = 0.6846$
$\mu_6 = 0.0000$
$\mu_7 = 0.0000$
$\mu_8 = 0.0000$



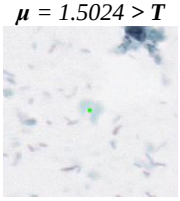
Decision: Non-Cell
$\boldsymbol{\mu}$ *= 2.3720 >* $\boldsymbol{T}$

$\mu_1 = 0.0000$
$\mu_2 = 0.0000$
$\mu_3 = 0.0000$
$\mu_4 = 1.0000$
$\mu_5 = 0.3817$
$\mu_6 = 0.5833$
$\mu_7 = 1.0000$
$\mu_8 = 0.0000$



Decision: Non-Cell
$\boldsymbol{\mu}$ *= 1.5024 >* $\boldsymbol{T}$

$\mu_1 = 0.0000$
$\mu_2 = 0.0000$
$\mu_3 = 1.0000$
$\mu_4 = 0.0000$
$\mu_5 = 0.5447$
$\mu_6 = 0.0833$
$\mu_7 = 0.0000$
$\mu_8 = 0.0000$



Decision: Non-Cell
$\boldsymbol{\mu}$ *= 3.5817 >* $\boldsymbol{T}$

$\mu_1 = 1.0000$
$\mu_2 = 1.0000$
$\mu_3 = 1.0000$
$\mu_4 = 0.0000$
$\mu_5 = 0.0000$
$\mu_6 = 0.0000$
$\mu_7 = 0.0000$
$\mu_8 = 0.7272$

Fig. 9. Examples of outputs for several candidates

The segmentation method presented in this paper has a number of adjustable parameters (the number of angles and distances used in re-dimensioning, the scaling factor for the fourth dimension of pixels, the type of distance for the adaptive k-means algorithm, merging rules, etc.) and one perspective would be to study the influence of these parameters on the overall classification. The resulting segmented image can be used to extract indicators whether the candidate is actually a cell or merely a non-cell.

Although we have used fuzzy classes for the final decision, the modularity of the implementation allows for a variety of methods to be used in decision making. Furthermore, the number of indicators may be increased and finer classifications may be attempted.

ACKNOWLEDGMENT

This work was fully supported by the Joint Applied Research Projects Intelligent System for Automatic Assistance of Cervical Cancer Diagnosis, grant number: PN-II-PT-PCCA-2013-4-0202, funded by Executive Unit for Higher Education, Research, Development and Innovation Funding (UEFISCDI).

REFERENCES

[1] G. N. Papanicolaou. A new procedure for staining vaginal smears. *Science, 95(2469): 438 – 439, 1942*

[2] A. Gençtav, S. Aksoy and S. Onder. Unsupervised segmentation and classification of cervical cell images. *Pattern Recognition, 45(12): 4151 – 4168, 2012.*

[3] Z. Lu, G. Carneiro and A. P. Bradley. Automated nucleus and cytoplasm segmentation of overlapping cervical cells. *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013, 8149: 452 – 460, 2013.*

[4] M. E. Plissiti, C. Nikou and A. Charchanti. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine, 15(2): 233 – 241, 2011.*

[5] R. Moshavegh, B. E. Bejnordi, A. Mehnert, K. Sujathan, P. Malm and E. Bengtsson. Automated segmentation of free-lying cell nuclei in pap smears for malignancy-associated change analysis. in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE: 5372 – 5375, 2012.*

[6] K. Li, Z. Lu, W. Liu and J. Yin. Cytoplasm and nucleus segmentation in cervical smear images using radiating GVF snake. *Pattern Recognition, 45(4): 1255 – 1264, 2012.*

[7] S. F. Yang-Mao, Y. K. Chan and Y. P. Chu. Edge enhancement nucleus and cytoplast contour detector of cervical smear images. *IEEE Transactions on Systems, Man and Cybernetics, 38(2): 353 – 366, 2008.*

[8] J. Angulo. A mathematical morphology approach to cell shape analysis. In *Progress in Industrial Mathematics at ECMI 2006, 543 – 547, 2008.*

[9] M. S. Nosrati and G. Hamarneh. Segmentation of overlapping cervical cells: A variational method with star-shape prior. *IEEE 12th International Symposium on Biomedical Imaging (ISBI), 186 – 189, 2015.*

[10] C. Rasche, S. Oprisescu, A. Sultana, T. Radulescu. Analysis of pap smear images with iso- and edge-contours. In: *IEEE 11th International Conference on Intelligent Computer Communication and Processing, 375 – 378, 2015.*

[11] P. Rousseuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics, 20: 53–65, 1987*