ISSN 2464-4617 (print) WSCG 2016 - 24th Conference on Computer Graphics, Visualization and Computer Vision 2016
ISSN 2464-4625 (CD-ROM)

Cytological Low-Quality Image Segmentation Using Nonlinear Regression, K-means and Watershed

Ramon A. S. Franco, Paulo S. Martins, Marco A.G. de Carvalho
University of Campinas (UNICAMP)
Brazil , 13484-332, Limeira, SP
{ramon,paulo,magic}@ft.unicamp.br

ABSTRACT

Since 1950, conventional cytology uses glass slides for microscopic analysis of cervical cells, in order to perform Pap Test. Such method yields low-quality images and overlapping cells, which both hampers their analysis and classification. Several countries use a modern method for the realization of Pap test called ThinPrep because it offers high- quality images and overcomes the problem of overlapping cells. ThinPrep facilitated the development of advanced image processing techniques for segmentation and classification of cervical cells. However, this method is not used by most of the developing countries of the world due to its relative high cost. This paper presents an algorithm for segmenting digital images obtained from conventional cytology method on glass slides. The technique uses Watershed Transform and K-Means Clustering in order to find cell markers or seeds. Nonlinear regression is applied as a way to refine the markers and to allow again the Watershed Transform utilization. We apply the technique in 10 glass slides of pap smears with a total of 67 cells. Our proposed technique has a promising performance in terms of accuracy of about 85%.

Keywords

Pap test, Image analysis, Watershed, K-means

1 INTRODUCTION

Cervical cancer remains the second leading cancer affecting women in developing countries [Glo01a]. The Pap Test, also known as oncological cytology or exfoliative cytology, has been used for the collection of human papillomavirus in the battle (i.e. prevention and diagnosis) against cervical cancer since 1950.

The Pap Test is a method developed by the physician George Papanicolaou for identifying neoplastic malignant or pre-malignant cells that precede the development of cancer [Mor01a]. This technique was originally developed to prevent cervical cancer. The cells are harvested in the region of the external orifice and endocervical canal, placed in a transparent glass slide, stained and taken for examination under the microscope.

Trained personnel can distinguish normal cells from malignant cells, such as those with indications of precancerous lesions [Nai01a]. However, the conventional Pap test technique has its limitations.

One of its limitations is the occurrence of false negatives (FN), i.e. when abnormal cells present in the test remain undetected. Aspects such as the manual spreading of cells leading to cell fragmentation on the glass slides is one of the reasons why false negatives occur in the Pap test. This normally occurs after obtaining cervical cells, where the physician must transfer cells using the spatula against the surface of the glass slides. The remaining cell assembly is disposed on the lamina, possibly leading to overlapped cells. If an abnormal cell is hidden under a healthy one, the pathologist would have difficulty locating them. Fig. 1 shows the appearance of a layer of cells collected by this method [Bud01a].

As we can see in Fig. 1, the cellular material is spread over the entire area, resulting from the friction of the wooden spatula on the laminated surface. To overcome the problem of overlapping cells of the cervix, a more efficient technology called ThinPrep was developed.

ThinPrep is a technology that mechanically separates the cervical cells by centrifugation, providing visibility of the cells and avoiding overlapping of cells [Loz01a]. In order to use the ThinPrep technology, it is necessary to change the conventional collection of glass by another collection called LCB [Ans01a]. Despite the advantages of the ThinPrep technology, it is not used by most countries due to its prohibitive costs and the lack of infrastructure needed for its implementation [Ama01a].

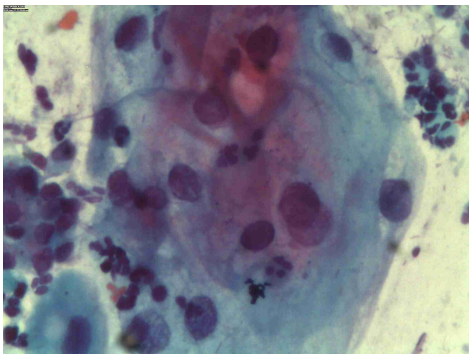

Figure 1: Glass slide of conventional Pap test

Much of recent research on image segmentation and image classification is founded on the analysis of images captured by the ThinPrep method, which considerably facilitates the task of pre-processing.

The goal of this paper is to segment images of cervical cells of low quality, obtained from conventional cytology. We apply clustering techniques and Watershed. A proposed solution is presented to improve the quality of segmentation. The proposed method combines K-means and Watershed in order to identify cervical cells, and it uses nonlinear regression for preprocessing of the histogram.

This paper is organized as follows: Section 2 presents a brief review of early work. The approach adopted is presented in Section 3. Section 4 shows initial results and discussions. Finally, in Section 5, we present the main conclusions and future work.

2 RELATED WORK

A nucleus and cytoplasm contour detector (NCC Detector) is presented by Pai et al. to automatically detect the cytoplasm and nucleus contours of a cell in a cervical smear image [Pai01a]. This detector has two different phases, according to the object to be segmented. In the cytoplasm detector, ATD method (a thresholding method) is used in order to draw the cytoplasm contour. The nucleus detector phase is composed of three stages: gradient calculation, the maximal gray-level-gradient-difference (MGLGD) method (proposed to sever the nucleus from the cytoplasm) and the contour connection. The experiments were accomplished with 50 cervical smear gray-level images, of 128x128 pixels. The results show that the NCC Detector is superior to two existing methods, the gradient vector flow-active contour model [ChaV01a] and the edge enhancement nucleus and cytoplasm contour detector.

Yung-Fu et al. proposed a method based on five steps well defined: image acquisition and categorization, image editing and processing, morphometry (contour segmentation of cell nucleus and cytoplasm, measurement and analysis), Support Vector Machine (SVM) classification and assessment of diagnostic performance[Yun01a]. The main goal is classifying four different types of cells and to discriminate dysplastic from normal cells. Besides, two experiments were conducted to verify the classification performance and results showed that average accuracies about 97%. The authors used a set of performance metrics and presented a good number of tables that shows the classification and diagnostic performance.

A method for automatic cervical cancer cell segmentation and classification was proposed by Chankong et al. [Cha01a]. His method provides a cell segmentation that separates nucleus, cytoplasm and background. The approach uses Fuzzy C-means (FCM) clustering technique and the results achieved a segmentation accuracy around 95%, according to the chosen classifier. The experiments have not included full-glass slides segmentation, i.e., the authors work with selected regions. The segmentation and classification performances were compared with C-means clustering and Watershed techniques. The comparison analysis showed that the proposed approach results in good performance and is better than the cited work.

Ushizima et al. [Ush01a] developed algorithms for quantitative analysis and pattern recognition from 2D cervical cells images. They proposed a pre-processing step, based on adaptive histogram equalization and mean shift technique, in order to make homogeneous the image regions and have an image with good contrast. The overall cytoplasm segmentation is accomplished using Watershed Transform. Before that, it is implemented some tasks in order to find the Watershed seeds, including clustering and hole filing. The authors provide a set of automated tools capable of detecting multiple cells obtained from ThinPrep Pap Test, including ROI (region of interest) selection, noise minimization and cell classification. A difficulty of the proposed method consists on the cytoplasm segmentation, according to the paper, generating a low accuracy.

In essence, our work differs from previous work by the one or more of the following features:

- *Low-quality images*: The set of images adopted in this work, as mentioned before, were obtained directly from relatively low-resolution cameras and standard microscopes, which is the reality of many developing countries; this contrast with much of the work which uses ThinPrep as the basis for the segmentation procedure;

- *Conventional testing*: The procedure employed is the conventional testing which uses manipulation of cells on a glass slide. The challenge with this process is the resulting undesired elements in the glass

slide, such as cell fragments, which may negatively interfere with the segmentation process;

- *Watershed transform*: We used the Watershed transform, which is arguably a satisfactory method for finding non-uniform contours such as the ones from cells. Clearly, there are various methods that can be used to segment images, or even a combination of methods. Ongoing work is experimenting with such variations; Whereas there has been a relative large number of work employing the Watershed, we are not aware of other projects dealing with cytological segmentation that uses Watershed twice or in a chained fashion, whereby the output of the first is the input to the second (Section 3);

- *Nonlinear regression:* Nonlinear regression, despite its relative simplicity and low computational overheads, proved to be an effective method for clearing the background image, and thus providing a much easier path for the ensuing segmentation process. Most work applies Otsu [Liu01a] in this stage, which we applied in an earlier stage of this work. However, the results obtained with the nonlinear regression filter were far superior, and therefore it was the method of our choice for this work.

Furthermore, it is not the focus of this work (for now) to experiment with classification of cells; our short term goals are to increase and maximize the accuracy of the method by exploring new variants to the proposed approach.

3 PROPOSED APPROACH

In this section we describe the approach employed to improve the quality of segmentation in low-quality images. It consists of 11 steps, which are illustrated in Fig. 2 and described through the remainder of this section.

To facilitate the discussion, these steps are grouped in three major classes, i.e. 1) Acquisition and categorization, 2) Image processing and 3) Cell Segmentation.

Acquisition and Categorization

The nucleous and cell morphology was used to calculate the ground truth in each image. The ratio of the cell area NC_r is given by:

$$NC_r = \frac{N_a}{N_a + N_c} \tag{1}$$

where N_a is the nucleus area and N_c is the cytoplasm area.

Image Preprocessing

Considering that images were of poor resolution, with non-homogeneous contrast and brightness, it was necessary to perform filtering on the image. Three filters were applied:

- *Bilateral filtering:* We applied the following nonlinear bilateral filter to preserve image energy and the image contours while simultaneously reducing noise:

$$I_f(x) = \frac{1}{W_p} \sum_{x_i \varepsilon \Omega} I(x_i) f_r(||I(x_i) - I(x)||) g_s(||x_i - x||) \tag{2}$$

  where: I_f is the filtered image; I is the original input image to be filtered; x are the coordinates of the current pixel being filtered; Ω is the window centered in x; f_r is the range kernel for smoothing differences in intensities; g_s is the spatial kernel for smoothing differences in coordinates. The range parameter σ_r was set to 9. The spatial parameter σ_d was set to 75.

- *Median filtering:* The median filter was applied with a kernel 3x3, in order to remove noise while preserving the edges and other details;

- *Unweighted average:* All images show a significant background representing the glass slide. The cytoplasm scattered across the glass slide and the undesired background in Fig. 3 would hinder the identification of tonal groups by the clustering algorithm, if not removed.

  Therefore, we decided to apply a nonlinear moving average filter in order to remove the background of the image (i.e. glass slide). We applied the equation MA = P_m + P_{m-1} +.. .+ $P_{m(n-1)}$)/ N where P_m is the average amount N and the number of samples [YaL01a].

  This nonlinear filter smoothens the histogram creating a harmonic function and enabling the determination of the maximum peaks. Consequently, the background image is removed as illustrated in Fig. 4. Thus, it was possible to automatically establish the threshold of each one of the glass slides image. Fig. 5 illustrates the image histogram before we applied the unweighted average filter, and Fig. 6 the histogram after the filtering.

  Clearly, the peak on the right side of the graph mainly belongs to a set of data pertaining to the bottom of the glass slide. These pixels were then removed before starting the image processing step.

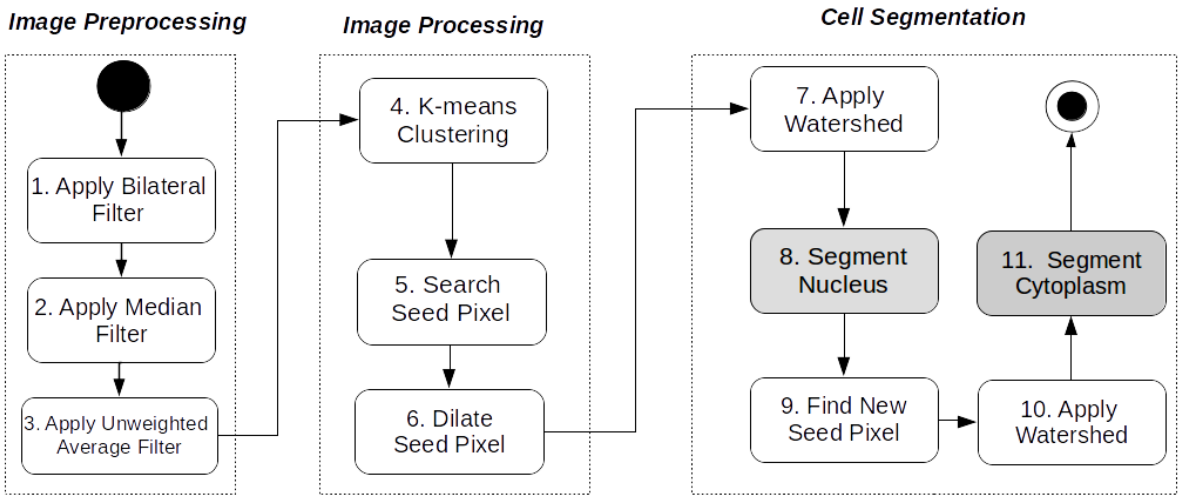

Figure 2: Activity Diagram of the Segmentation Process

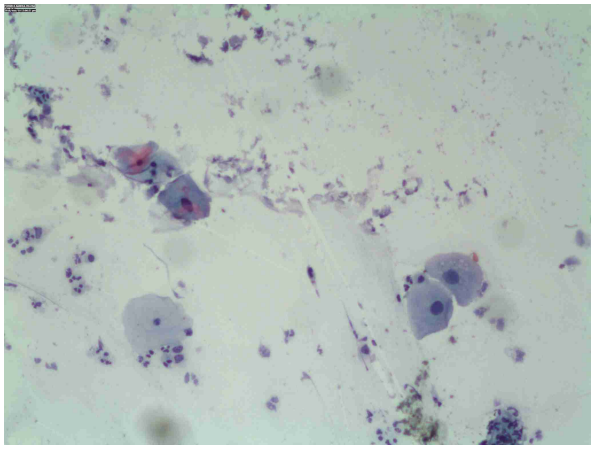

Figure 3: Low-quality image of cervical cells in glass slide

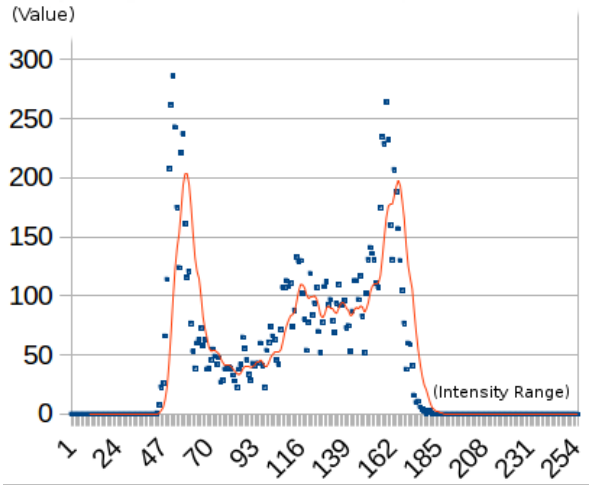

Figure 5: Original histogram (pixel frequency vs gray-level value)

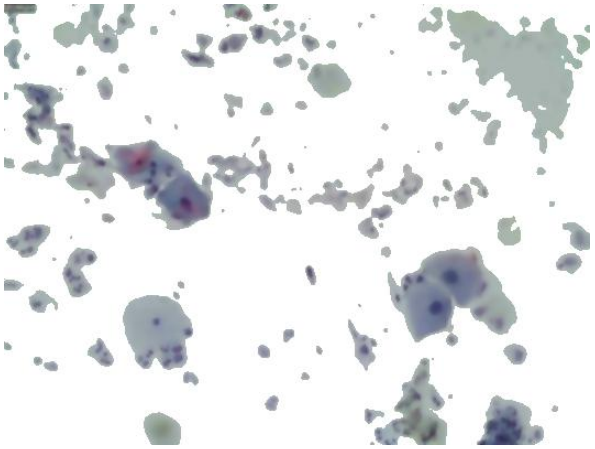

Figure 4: Isolating the cells from the glass slide

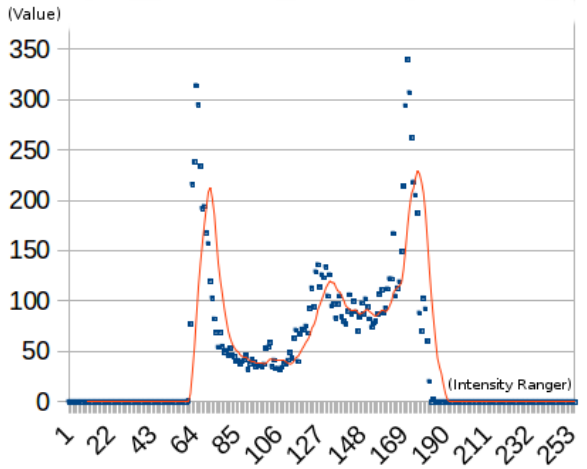

Figure 6: Histogram modified by the unweighted moving average filter

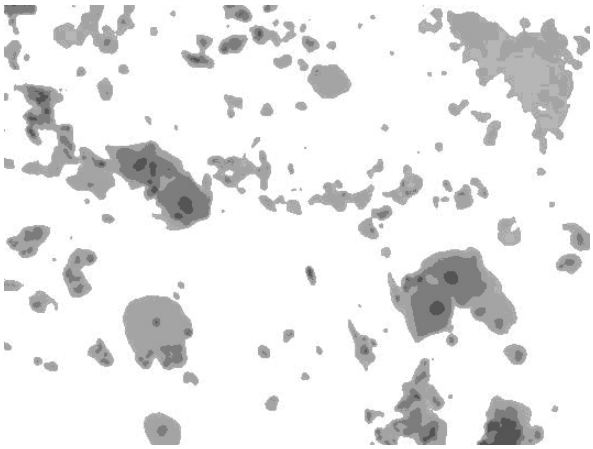

Figure 7: The image after clustering

Image Processing

After preprocessing the images, the K-Means clustering algorithm is applied in order to create distinct gray-scale groups [Har01a]. The goal of using K-Means is exclusively to find distinct gray-scale levels of seed pixels for the segmentation using the Watershed algorithm [Beu01a]. The number of groups selected was seven, as this number had shown the best segmentation results in a number of experiments varying the number of groups. Fig. 7 shows the results observed after image clustering.

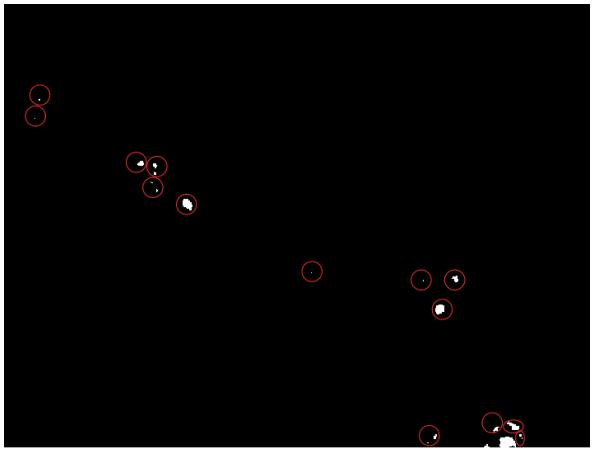

Figure 8: Pixels seed

After clustering, the cytoplasm still does not show well since it is illustrated within four different shades of gray (i.e. groups of gray levels). Note that the background is one among the four groups. Note also that this technique also detects other elements such as cell fragments, resulting from the mechanical handling of the samples on the glass slide. On the other hand, the nuclei has successfully been isolated in the image, i.e. within the darkest areas. Thus, a search algorithm was used to determine the two most prominent (darkest) groups, as such groups are the ones that capture the cell nuclei.

The pixels at the coordinates output by the clustering process were then dilated with a scale factor 3 to facilitate their identification as seed pixels, and thereby the most significant nuclei. These pixels were used in the following step in the Watershed algorithm, to help determination of the masks and the unknown area of the image. Fig. 8 shows some of the resulting pixels seeds.

Cell Segmentation

Cell segmentation was carried out in two steps, each consisting of one application of the Watershed. This process ensured the proper segmentation o the cytoplasm:

- *Watershed 1*: The goal of this first Watershed step was to identify the nuclei boundaries. The seed pixels were the ones found earlier by the clustering process. A good way to understand the Watershed is by comparing its operation to that of a flooding of a water basin. In the Watershed method, an image is segmented by constructing the catchment basins, or lakes, of the image. The image is flooded starting from the seeds, where each seed correspond to a lake, until the whole image has been flooded. A dam is built between lakes that meet with others lakes. At the end of flooding process, we obtain one region for each catchment basin of the image [Beu01a].
- *Watershed 2*: In a second step, subsequently, we reapplied the Watershed algorithm. However, this time using as the new seed pixels the nuclei detected above. In this case the goal is to find the cytoplasm boundaries.

The K-Means Clustering algorithm was able to find the seed pixels that were subsequently used by the first implementation of the Watershed Transform (Watershed 1, as cited above

4 RESULTS

In this section we show a comparison analysis of the results obtained by manual segmentation carried out by the domain expert (i.e. pathologist) against the segmentation obtained by the automatic segmentation approach. In essence, the area of the cytoplasm and the nucleous found by the pathologist were compared to the respective areas found by the automatic segmentation. This procedure was repeated throughout 10 glass slides.

The details of this analysis are shown in Table 1, where A_{cp} is the average area of cytoplasm identified by the pathologist (manual segmentation); A_{np} is the average area of nuclei identified by the pathologist (manual segmentation); A_{cs} is the average area of cytoplasm identified by the segmentation algorithm (automatic segmentation), and A_{ns} is the average area of nuclei identified

Glass Slide #	# Cells	A_{cp}	A_{np}	A_{cs}	A_{ns}	Cytoplasm error (%)	Nuclei error (%)	Accuracy (%)
1	6	10551.63	558.88	11110.51	730.51	5.03	23.50	81.53
2	5	63934	5285.99	58798.46	5536.57	8.73	4.53	95.79
3	14	16823.85	1041.67	17328.23	1431.19	2.91	27.22	75.69
4	1	1237.81	76.74	1348.14	137.5	8.18	44.19	63.99
5	7	18061.66	1118.41	18676.37	1568.69	11.71	19.52	68.67
6	5	19299.47	1195.15	20024.51	1706.19	6.12	24.08	82.04
7	6	4485.46	459.83	4448.96	699.65	0.81	34.28	66.54
8	8	6157.05	784.5	4788.19	748.96	22.23	0.74	78.51
9	6	29883.9	868.93	29816.27	742.56	0.23	14.54	76.15
10	9	6262.5	1086.11	6834.34	1184.5	8.37	8.31	74.09
Results	67	176697.33	12476.21	273174	14486.32	7.43	20.09	85.00

Table 1: Accuracy of the segmentation (Acp and Anp, manual segmentations; Acs and Ans, automatic segmentations)

by the segmentation algorithm (automatic segmentation). For example, in glass slide 7 six cells were identified; the total area of cytoplasm and the total area of nuclei are shown for both manual and computational procedures; the percentage cytoplasm error is 0.81, and the corresponding percentage nuclei error is 34.28; the accuracy for this slide is 66.5%. Considering all glass slides, the total accuracy was 85 %.

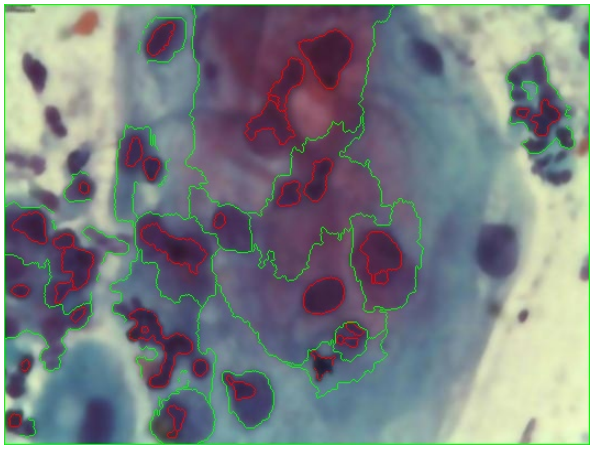

Figure 9: Final segmentation (glass slide 3)

As shown in Table 1, the segmentation accuracy was 85%. The method segmented 60% of the cells marked by the pathologist. On the other hand, it was able to segment 20% of the cells that were not identified by the pathologist.

Fig. 9 illustrates the final segmentation of the cytological low-quality image (glass slide 3) using nonlinear regression, K-Means clustering and Watershed Transform. As we can see, the algorithm segmented 15 cytoplasms and 28 nuclei, which were scattered throughout the image. 30% of the nuclei identified through automatic, image-processing segmentation, were not easily identifiable by pathologists because of the poor quality of the images.

5 CONCLUSION

This work is intended to improve the availability of current image processing technologies in countries and areas that are not able to afford modern collection methods such as Thin Prep.

The use of glass slide for conventional Pap test cytologies is quite common in developing countries, due to their low costs and easy implementation. However, the analysis of generated samples by image processing algorithms is a challenge: many of the images collected on glass slide are not exploited on account of their low quality and visibility. Without a method that is conceived to address such concerns, pathologists are not able to perform diagnostics using these images. This leads to extra losses, since the samples cannot be processed and have to be simply eliminated.

We performed image segmentation using images of poor quality, by means of nonlinear regression, K- means clustering and Watershed transform. We conducted experiments with 10 glass slides containing 67 cells previously measured by the pathologists. In addition, we achieved cell nuclei segmentation, which were not labeled by the pathologist by low visibility.

The segmentation accuracy was 85%. Considering that such images are currently discarded by the pathologist because of their low quality, this may already be deemed an acceptable result. However, in future work, we intend to carry out further experiments using complementary techniques to improve the accuracy as well as the overall citoplama's segmentation.

6 ACKNOWLEDGEMENTS

The authors would like to thank Synergy Telemedicine Company S.A. from Colombia for allowing the use of Pap smear slide images for investigation purposes. Ramon Franco is supported by Brazilian Agency CAPES - (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

7 REFERENCES

[Mor01a] Morrell, S., Taylor, R. and Wain, G. A study of Pap test history and histologically determined cervical cancer in NSW women, 1997-2003., J. Medical Screen., vol. 12, no. 4, pp. 190-196, 2005.

[Nai01a] Naib, Z. M. Pap Test, in Clinical Methods: The History, Physical, and Laboratory Examinations, 1990.

[Bud01a] Budge, M., Halford, J., Haran, M., Mein, J. and Wright, G. Comparison of a self-administered tampon ThinPrep test with conventional pap smears for cervical cytology, Aust. New Zeal. J. Obstet. Gynaecol., vol. 45, no. 3, pp. 215-219, 2005.

[Loz01a] Lozano, M., De Miguel, C. and Cousillas, Nuevas tecnologias en Citopatologia, in Sociedad Espanola de Anatomia Patologica, p.143, 2011.

[Ans01a] Anschau, F., Goncalves, M. A. G. Citologia Cervical em Meio Liquido Versus Citologia Convencional, Femina, vol. 34, no. 05, pp. 329-335, 2006.

[Ama01a] Amaya, M., Acosta, P., Mora, M. Citología convencional y en base liquida en muestra compartida de tomas cervocouterinas, Reperto Medico y Cirugia., vol. 24, no. April, pp. 41-47, 2016.

[Glo01a] GLOBOCAN. Estimate Cancer Incidence, Mortality and prevalence Wordwde in 2012, 2012. [Online]. Available: http://globocan.iarc.fr/Default.aspx.

[Liu01a] Liu, D. and Yu, J. Otsu method and K-means, in Proceedings - 2009 9th International Conference on Hybrid Intelligent Systems, HIS 2009, 2009, vol. 1, pp. 344-349.

[Pai01a] Pai, C. Chang, and Y. K. Chan, Nucleus and cytoplast contour detector from a cervical smear image, Expert Syst. Appl., vol. 39, no. 1, pp. 154-161, 2012.

[Yun01a] Yung-Fu, C. Semi-Automatic Segmentation and Classification of Pap Smear Cells, IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 1, 2014.

[Cha01a] Chankong, T., Theera-Umpon, N. and Auephanwiriyakul, S. Automatic cervical cell segmentation and classification in Pap smears, Computer Methods and Programs in Biomedicine, vol. 113, no. 2, pp. 539-556, 2014.

[Ush01a] Ushizima, D. M., Gomes, A. H., Carneiro, C. M. and Bianchi, A. G. C. Automated Pap Smear Cell Analysis: Optimizing the Cervix Cytological Examination, in 2013 12th International Conference on Machine Learning and Applications, pp. 441-444, 2013.

[Yal01a] Ya-Lun. Chou, Statistical Analysis With Business and Economic Applications, 2nd ed. Mishawaka, 1975.

[Har01a] Hartigan, J. A. and M. A. Wong, A K-Means Clustering Algorithm, Appl. Stat., vol. 28, no. 1, pp. 100-108, 1979.

[Beu01a] Beucher, S. The Watershed Transformation Applied to Image Segmentation, in Proceedings of the 10th Pfefferkorn Conference on Signal and Image Processing in Microscopy and Microanalysis, 1992, pp. 299-314.

[ChaV01a] Chan, T. F., and Vese, L. A. Active contours without edges. IEEE Transactions on Image Processing, 10(2), 266-277, 2001.

[LiY01a] Liu, D. and Yu, J. Otsu method and K-means,Ó in Proceedings of the 9th International Conference on Hybrid Intelligent Systems, vol. 1, pp. 344-349, 2009.