2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Automatic Epithelial Cells Detection of Pap smears images using Fuzzy C-Means Clustering

Izzati Muhimmah [1 +], Rahadian Kurniawan [1] and Indrayanti [2]

[1] Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
[2] Faculty of Medicine, Universitas Muhammadiyah Yogyakarta, Indonesia

Abstract. The task of epithelial cells counting of pap smears images is still a burden for the pathologists. The computer-assisted system to address this challenge will benefit them. However, identification of epithelial cells is nontrivial task due to the complexity nature of pap smears data. This article proposes combination of distance-metric and Fuzzy C-Means clustering to identify the nuclei of epitheleal cells for any given pap smears images. The proposed method is evaluated using Liquid Based Preparation/ ThinPrep, which contain 350 epithelial cells which are confirmed by our expert Pathologist. The proposed methodology successfully automatically identify of cells with sensitivity rate of 90.86% and specificity rate of 81.62% in comparison to the expert assessment on NCI Bethesda Systems of pap smears data.

Keywords: pap smears, epithelial cells, nuclei detection, FCM

1. Introduction

The challenge of pap smears slide reading is known to be very time-consuming and tedious. It is due to the characteristics of pap smears images which suffer from low contrast, in homogeneities (cells tend not to absorbed the stained materials equally), and cluttered (uneven spreading of cells on slide) [1]. Often in practical setting, the task of counting the number of epithelial cells under microscope to measure its adequacy was not done because it is very difficult to do manually.

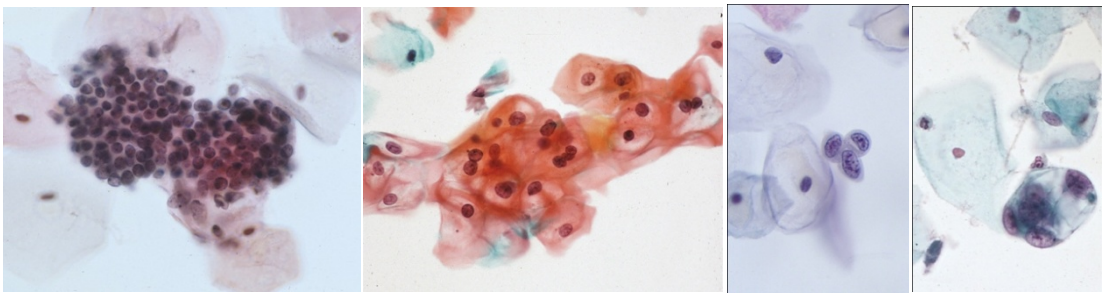

Fig. 1: Examples of pap smears images. Note that the images broadly varied w.r.t. magnifications and color intensities.

Pap smears images are varied with respect to magnification, contrast agent, cell types. These broad variations can be seen in Figure 1. As consequent, automatic detection using fixed parameters cannot be done. In other words, predefined thresholds on color intensities or imposing a single value of cells metric may not work.

There are some previous works to address this challenge by mean of automatic cells detection. Several approaches have been proposed such as: used deformable templates to identify cells with applied a generalized Hough transform to roughly detect round-like shapes [1], used a new criterion function based on

+ Corresponding author. Tel.: + 62274895287; fax: +62274895007.
E-mail address: izzati@uii.ac.id.

statistical structure of the objects in cell image [9], automatically finding the threshold value and detect the edges of cytoplasm and nucleus edges using moving k-means and Modified Seed Based Region Growing [10]. Although these methods present promising results, they were still using grayscale image and not yet tested on RGB images. Thus, this paper investigates automatic method for detection of epithelial cell location in RGB Pap Smear images.

The reminder of this paper is outlined as follows: the proposed methods for epithelial cells detection are described in Section 2. The data and the evaluation methods are explained in Section 3. Section 4 shows our results and discussions on our findings including pointing out the future works

2. Epithelial cells detection

2.1. Preprocessing

As mentioned earlier, the Pap smears images may suffer from low contrast. So, the preprocessing step is needed in order to tackle this. In this step we adopted method proposed by Plissiti, et al [2] which can be summarized as follows.

In a first step, we perform the contrast limited adaptive histogram equalization [3] and global thresholding using the method proposed by Otsu [4] to the red, green and blue channels of the initial image into three binary images. Subsequently, a binary mask is obtained as the result of a logical OR operation of those three binary images. The detected areas of the binary mask are extended with a morphological dilation. Then, all the connected components with an area smaller than 500 pixels are removed. This is necessary for the exclusion of image artifacts that may interfere in the next steps.

This pre-processing stage allows us to have regions of interest (ROI) removed from its background. Yet, the ROI may consist of the targeted-nuclei and cytoplasm. Thus, further step is needed in order to detect the nuclei only.

2.2. Candidate of Nuclei Detection

One can identify nuclei from cytoplasm is from its darker color (low intensity) than the surrounding Cytoplasm tends to have lighter color (high intensity). For this reason we search for intensity valleys using the h-minima transform [5] in the red, green and blue channels of the outcome preprocessing image. Next, we subtract a threshold h from every pixel of the outcome image. A morphological reconstruction [6] step is then performed. In this step, we use the resulted image as a marker and the initial image as a mask. Then, we subtract the red, green and blue channels of the outcome preprocessing image with the outcome image from morphological reconstruction step.

However, the resulting images of the region-minima approach may have several markers within the targeted-nuclei. Also, there are some markers on the cytoplasm. So, edge is the best feature to impose selection only within the nuclei boundary.

Furthermore, the number of markers are reduced using a combination of Euclidean distance measure and their intensities similarity. Two markers which have shorter distances will be compared their colour intensities. Markers with low intensity will survive, whereas markers with high intensity will be lost. Illustration of this stage can be seen in Figure 2.

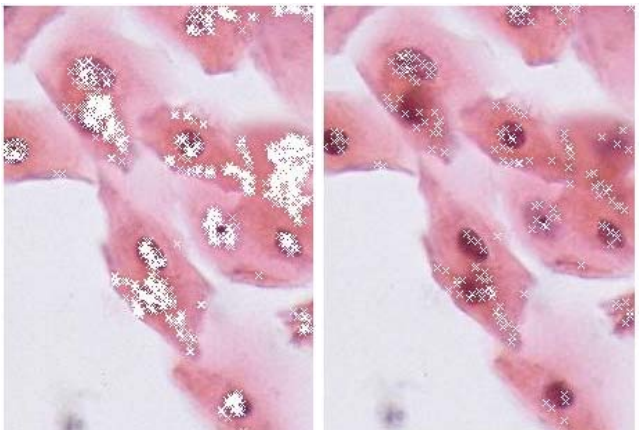

Fig. 2: nuclei detection and refinement.

2.3. Refinement of Candidate Cell Nuclei Centroids

Centroids that have been detected in the previous stage will be reduced by performing morphological dilation at each pixel where centroids are located. Next, we look for pixels with the lowest intensity as new centroids location of the area formed by the algorithm. For all the obtained centroids we apply the following rule:

```
i =1
Repeat
    ∀p = (x, y) ∈ LA(i)
    Select r = {p | min(I(p))}
    i++
Until i=nL
```

LA is an area that has a label equal to *i*, *nL* is the number of labeled area, *I* is a grayscale image of the original image.

2.4. Reduction in Number of Centroids by Distance-Metric

The number of markers are reduced using a combination of Euclidean distance-metric and their intensities similarity. Two markers which have shorter distances will be compared their color intensities. Markers with low intensity will survive, whereas markers with high intensity will be lost.

2.5. Clustering

Despite of the vast markers reduction on the candidate nuclei detection stage, there are remain some markers within nuclei or on cytoplasm. We aimed to get one marker to represent one nuclei. To achieve that, we proposed the Fuzzy C-Means Clustering as follows:

1. Two clusters are defined, one represents the nuclei and another belongs to not nuclei.
2. The classification criteria are the closest distance between markers and similarity on their intensity values.
3. Do min-max normalization on both these criteria.
4. A positive class is defined when the average intensity of the cluster is lower than the other.

This proposed algorithm is effective to identify the nuclei in one-to-one relation. The result of this stage can be seen in Figure 3.

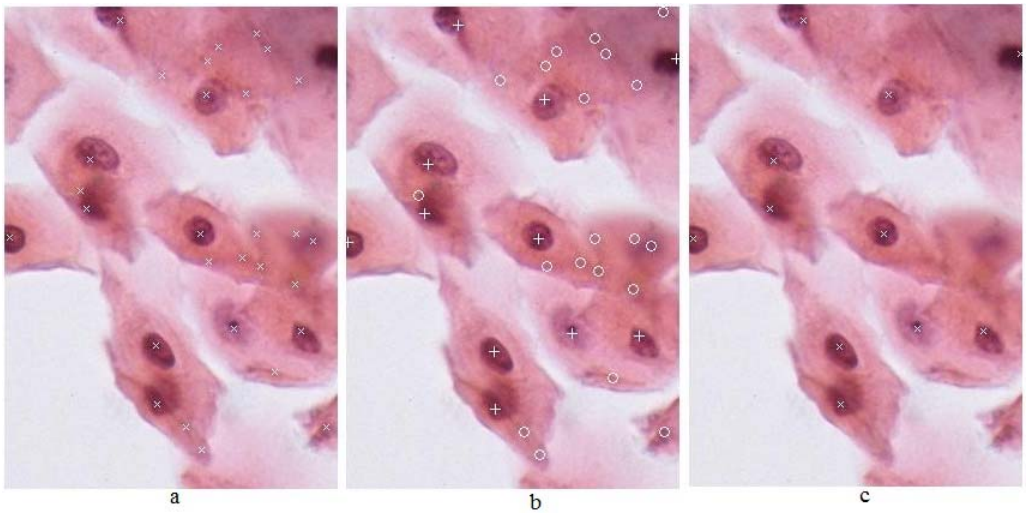

Fig. 3: Clustering stages: (a) initial image as result of distance-metric, (b) result of FCM: (+) denotes positive nuclei and (○) denotes not nuclei, and (c) one-to-one map of nuclei.

3. Evaluation method

3.1. Data

We tested our proposed methods on 11 pap smears images of NCI Bethesda System [7]. These consist of five images of medium magnification, and six of high ones. These images have 350 nuclei and have been confirmed by our pathologist [8]. This number was obtained as consistent nuclei identification by our expert [8]. It should be noted that the two readings were done in random ordered and had eight days interval in between the two readings to minimize the recall bias by our expert pathologist.

3.2. Evaluation methods

Furthermore, as a measure of the computational efficiency of the segmentation method, we present in Table I the processing times of the individual steps of the method developed in MATLAB using a Pentium 2.66 GHz and 4 GB of RAM.

TABLE 1 EXECUTION TIME OF THE PROPOSED METHOD

Step of the proposed method	Time in sec.(mean ± std)
Preprocessing	3.65 ±1.40
Candidates of nuclei detection	20.56 ± 8.58
Refinement of Candidate Cell Nuclei Centroids	3.18 ± 1.35
Reduction in Number of Centroids by Distance-Metric	2.17 ± 3.03

4. Results and Discussion

We tested our proposed methods with respect to its detection rate. The detection rate is compared with the expert truth (Table II, second column) in terms of its specificity and sensitivity. The detection rates of the proposed method can be seen in Table II. The average sensitivity rate was 90.86% and specificity rate was 81.62% which are quite promising.

TABLE 2 RESULT OF THE PROPOSED METHOD

File (*.jpg)	Expert Truth	True positive	True negative	False positive	False negative	Sensitivity	Specificity
1982	100	96	72	28	4	96%	72%
3079	122	108	122	0	14	89%	100%
6676	21	21	21	4	0	100%	84%
7214	28	25	31	2	3	89%	94%
7808	8	8	25	16	0	100%	61%

9945	24	23	14	12	1	96%	54%
6451	11	5	13	7	6	45%	65%
2064	6	6	6	0	0	100%	100%
2905	8	5	12	4	3	63%	75%
3835	15	14	19	4	1	93%	83%
9299	7	7	7	0	0	100%	100%
Sum	350	318	342	77	32	90.86%	81.62%

As can be seen in Table II, file 6451 and 2905 had a very low sensitivity and specivity rate. This was mainly caused by inhomogenity of its intensities (see Figure 4 for illustration). The cytoplasms on this image were darker than its nuclei. We modelled our methodolgy with assumption that nuclei is darker than cytoplams. Although our model correctly marks the nuclei on candidate detection stage, but these markers were lost on clustering stage.

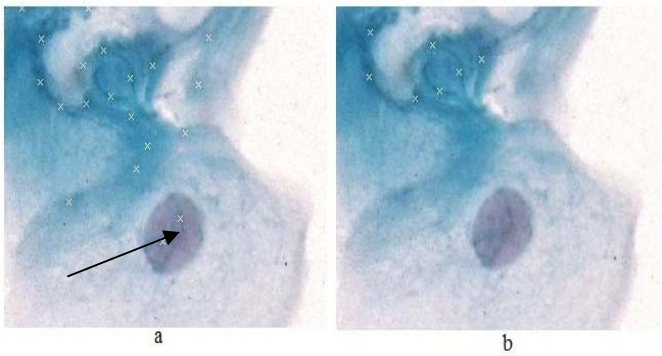

Fig. 4: False negative case. (a) Correct nuclei identification on candidate nuclei detection stage, but (b) lost on clustering stage.

On the other hand, there were cases where markers on the inflammatory cells (see arrow on Figure 5) rather than on the nuclei. This is caused by the intensity of the inflammatory cells were lower than those of the nuclei's.

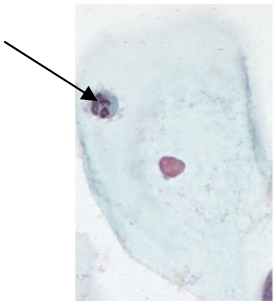

Fig. 5: Example of false positive case. The marker tends to stand on inflammatory cells rather than on the nuclei.

Despite our aim to obtain a one marker-to-one nuclei mapping, there are some nuclei that have several markers on. This is affected by variations on magnification values used in NCI Bethesda Systems [7]. Hence, further investigations need to be done to determine data-driven distance criteria for the clustering stage.

Our proposed methodology works quite well on identifying the epithelial cells in one-to-one mapping. This can serve as basis for further investigation on classifiying whether a marker stands on which type of epithelial cells. This will be the further avenue of our research.

5. References

[1] A. Garrido, N. Pérez de la Blanca, "Applying deformable templates for cell image segmentation", *Pattern Recognition*, vol. 33, 2000, pp. 821-832.

[2] ME. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of Cell Nuclei in Pap Smear Images Using

Morphological Reconstruction and Clustering," *IEEE Trans on Information technology in Biomedicine.*, vol. 15, no 2, March. 2011, pp. 233-241.

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

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

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

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

[7] NCI Bethesda System (http://nih.techriver.net/)

[8] R. Kurniawan, "Otomatisasi Deteksi Jumlah Nuclei pada Citra Pap Smear Menggunakan Kriteria Jarak & Clustering," M.S. thesis, Dept. Informatics., Universitas Islam Indonesia, Yogyakarta, Indonesia, 2011.

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

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