

Segmentation of Overlapping Cervical Cells in Normal Pap Smear Images Using Distance-Metric and Morphological Operation

Rahadian Kurniawan¹, Izzati Muhimmah², Arrie Kurniawardhani³, and Indrayanti⁴

^{1,2,3}Department of Informatics, University of Islam Indonesia, Yogyakarta 55501, Indonesia

Emails: ¹rahadiankurniawan@uii.ac.id, ²izzati@uii.ac.id, ³arrie.kurniawardhani@uii.ac.id

⁴Medical Faculty, Universitas Muhammadiyah Yogyakarta, Kasihan Bantul 55183, Indonesia

Abstract—The automatic interpretation of Pap Smear image is one of challenging issues in some aspects. Accurate segmentation for each cell is an important procedure that must be done so that no information is lost during the evaluation process. However, the presence of overlapping cells in Pap Smear image make the automated analysis of these cytology images become more difficult. In most of the studies, cytoplasm segmentation is the difficult stage because the boundaries between cells are very thin. In this study, we propose an algorithm that can segment the overlapping cytoplasm. First, the morphology operation and global thresholding to segment cytoplasm is done. Second, the overlapping area on cytoplasm region is separated using morphological operation and distance criteria on each pixel. The proposed method has been evaluated against the results of manual tracing by experts. The experiment results show that the proposed method can segment the overlapping cytoplasm as similar as experts do, i.e., 2.897 ± 3.632 (mean \pm std) using Hausdorff distance.

Index Terms—Cervical Cancer Cell, Overlapping Cells, Distance Metrics, Morphological Operation, Pap Smear

I. INTRODUCTION

PAP Smear is one of methods to examine the cervical cells. Those cells are examined by microscope to observe the alternation or acuteness of cervical epithelial cell as the initial sign of cancerous existence. Cervical cancer is the fourth highest cancer that attacks woman. There were more than 530.000 new cases of cervical cancer in 2012 [1]. Cervical cancer is a cancer arising from the cervix, the female reproductive system, which is in the upper part of the vagina and posterior portion of the uterus. Nowadays, in developing countries Pap Smear test used for examining cervical cells has been utilized extensively. It can decrease mortality rates caused by cervical cancer significantly. However, Pap Smear images are difficulties

to interpret. It is caused by clustered cells, overlapping cells, existence of inflammatory cells, blood stains, low contrast, and variation in illumination occurred because of inconsistent staining methods such as dye concentration. Figure 1(a) shows Pap Smear image that contains inflammatory cells, whereas Fig. 1(b) shows Pap Smear image that contains overlapping cells. Those difficulties consumes time and require a highly trained personnel to avoid high errors. Consequently, in recent years, there are a lot of researches focusing on building system that can interpret Pap Smear images automatically.

Currently, researches in the field of computer vision about cervical cancer have been carried out. This suggests that the need to combat this disease is very large. In the previous study [2], clustering by using the Morphological Reconstruction successfully detects most of nuclei in the image of the Pap Smear. A similar study to detect nuclei in the image of the Pap Smear is also performed using Fuzzy C-Means Clustering (CFM) [3]. The other research techniques that have been successfully used to segment nuclei in the Pap Smear images are deformable templates [4], pixel classification schemes [5], morphological operation and watershed transformation [6].

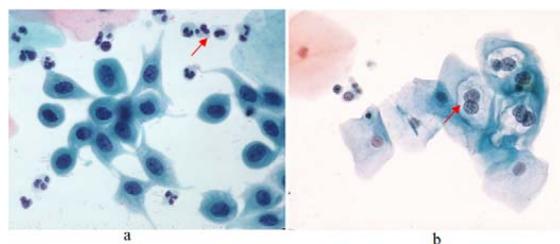


Fig. 1. (a) Pap Smear image that contains inflammatory cells and (b) Pap Smear image that contains overlapping cells. Source: (NCI Bethesda System).

Almost all of these studies only targeted the process of detection, analysis, and segmentation of nuclei regions by assuming that the accuracy of Pap Smear results is strongly influenced by the general appearance of nuclei on the image. The nuclei shows significant changes when the cell is infected by the disease. The detection of cytoplasm regions is also important because the cytoplasm features have proved very useful for the identification of abnormal cells. Cytoplasm is the cell membrane outside the nucleus. The studied conducted by Yang-Mao et al. [7] and Li et al. [8] had successfully performed segmentation of nuclei and cytoplasm area using Gradient Vector Flow (GVF) snake. However, the study only uses a single cell as an input to be analyzed. While each image preparation can contain up to 300 000 cervical cells [9] thus the analytical approach using a single cell cannot represent the actual case. Similar study that use multiple cell has been done by Ref. [10]. However, it has not overcome the cytoplasm overlapping. Figure 2 shows the dissimilarity between a Pap Smear image containing a single cell and a Pap Smear image containing multiple cells and overlapping cells.

This study is conducted to assist cervical epithelial cell identification automatically in Pap Smear images, namely segmentation of overlapping cervical cells. We focus on the nuclei and cytoplasm segmentation in Pap Smear image that contains multiple cells. Two main steps to segment the overlapping cervical cells proposed in this study are: nuclei segmentation and cytoplasm segmentation.

II. RESEARCH METHOD

A. Nuclei Segmentation

The nuclei are segmented automatically using the technique we proposed in Ref. [12]. Nuclei are segmented to obtain the actual boundaries of nuclei. Figure 3 illustrates the steps for nuclei segmentation.

In the first step, because Pap Smear images tend to have low contrast as shown in Fig. 4(a), preprocessing

is needed to increase the contrast between nuclei and background. Techniques used to increase the contrast are h -minima transform operation [12]. The h -minima transform operation is applied to each color space layer and subtraction operation between two images. Figure 4(b) shows the result of preprocessing. We can see that the contrast of image in Fig. 4(b) become higher. The nuclei segmentation process become easier and more precise when it is performed in high contrast image (Fig. 4(b)) than in low contrast image (Fig. 4(a)).

In the second step, the segmentation using morphological operation, and global thresholding operation approach is performed to identify nucleus candidate. The global thresholding operation used in this step are based on method proposed by Ref. [13]. Furthermore, modified watershed transformation proposed by Ref. [14] is used to address the overlapping nuclei. Figure 5(b) shows the result of the second step. All nucleus candidates that can be identified are marked by "x".

The last step is clustering by using FCM to separate nuclei against non-nuclei objects, such as inflammatory cells or blood stains. Inflammatory cells that exist almost in all Pap Smear images can disturb the automatic analysis process. Features that are used to separate the nuclei against non-nuclei object are the minor axis length, the average radius, the equivalent

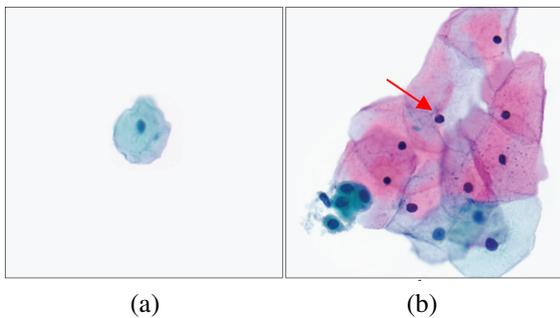


Fig. 2. Image comparison [11]. (a) Single cell, (b) multiple cells.

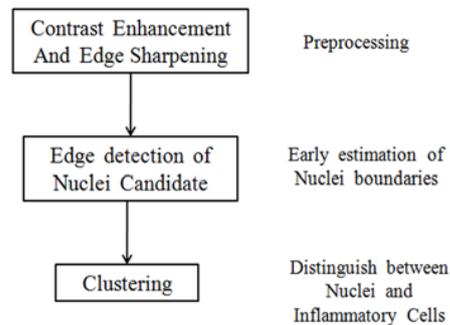


Fig. 3. The schemes for nuclei segmentation [3].

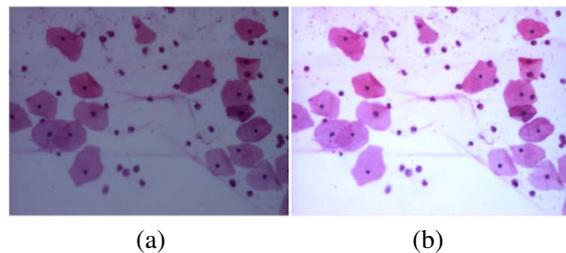


Fig. 4. (a) Original image, (b) Output image from preprocessing step.

diameter, the major axis length, the uniformity, the foreground-background contrast in red, the compactness, the circularity, and the eccentricity. Those nine features are used to identify nuclei based on Ref. [12]. In that study, those nine features have been tested in four different datasets and give the high accuracy for nuclei clustering. Figure 5(c) shows the result of the clustering process. Identified nuclei are marked by "+", whereas non-nuclei objects are marked by "o". Figure 5(d) shows the result of nuclei identification after eliminating non-nuclei objects.

B. Cytoplasm Segmentation

1) *Cytoplasm Boundaries Detection*: Cytoplasm segmentation is more difficult than nuclei segmentation. It is because nuclei intensities are darker than cytoplasm intensities. It is easier to distinguish nuclei against the background. On the contrary, cytoplasm tends to have similar intensity with background intensity. Hence, the first step of cytoplasm segmentation is a contrast enhancement using h -minima transform operation into each color space layer in order to make cytoplasm more homogenous. The values of h -minima at each color space layer (red, green, and blue) are calculated by

$$h_r = \frac{\mu(I_r)}{2}; \quad h_g = \frac{\mu(I_g)}{2}; \quad h_b = \frac{\mu(I_b)}{2}, \quad (1)$$

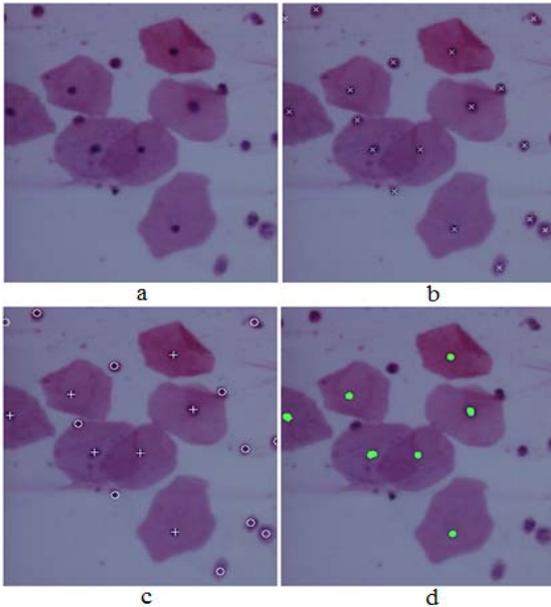


Fig. 5. (a) Original image, (b) Give "x" mark for nuclei candidate, (c) FCM Clustering results: "+" mark for nuclei cluster and "o" mark for non-nuclei object, (d) Nuclei Cluster.

where μ is the value of average intensities and I is the intensity value in the original image. Each layer of color that has been enhanced is converted to grayscale color space by using:

$$I_g = 0.2989h_r + 0.5870h_g + 0.1140h_b, \quad (2)$$

where h_r , h_g , and h_b are h -minima values at each color space layer (red, green, blue). After the grayscale images of h -minima transform image are obtained, the morphological opening operation is applied to the grayscale image:

$$I_{\text{opened}} = I_g \circ \text{SE} \quad (3)$$

where \circ denotes the opening operation [15]. The variable I_g is a grayscale image of h -minima transform image and SE is flat disk-shaped structuring element with radius of 5 pixels. The h -minima transformation with $h = 15$ is applied again to I_{opened} image in order to get more uniform color intensities of cytoplasm regions.

The output image of this stage is called H_i image that has more homogeneous color intensity in the cytoplasm regions.

The cytoplasm regions in that image are extracted by using:

$$C_n(p) = I_g(p) + H_i(p), \quad p \in D_I \quad (4)$$

where I_g is a grayscale transformation image obtained using Eq. (2), and H_i is h -minima transform image gotten from the previous process. Then, contrast enhancement filter is applied to complement the image of C_n in order to enhance the color intensities of an image.

The morphological dilation using a flat disk-shaped structuring element with radius of 3 pixels is applied by:

$$I_{\text{dilate}} = I_{\text{ce}} \oplus \text{SE}, \quad (5)$$

where \oplus denotes the dilation operation [15] and I_{ce} denotes the contrast enhancement filter image. Another grayscale image that will be used for marker image is obtained by:

$$I_{\text{mark}} = G - (I_{\text{dilate}} - I_g) \quad (6)$$

where G is a grayscale image of the original image, I_{dilate} is image gotten from Eq. (5), and I_g is image obtained from Eq. (2). After that, the binary image of I_{mark} is obtained using a thresholding process in each pixel of I_{mark} image. The thresholding process is given in Alg. 1 where D_I is the dimension size of image, and T is obtained from Eq. (7).

$$T = \frac{\mu(I_g)}{2} \quad (7)$$

Algorithm 1: The pseudo code for the thresholding process.

```

if ( $I_{mark}(p) < T, p \in D_I$ ) then
  |  $M(p) = 1$ 
else
  |  $M(p) = 0$ 
end

```

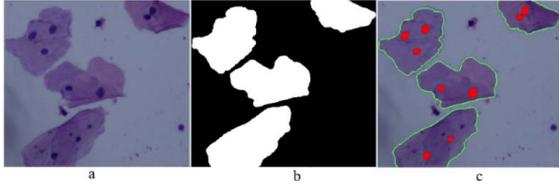


Fig. 6. (a) Original image, (b) Binary image from cytoplasm segmentation process, (c) Original image with boundary marker of cytoplasm.

After we get the marker image (M), a morphological opening is applied to M using:

$$BM = M \circ SE \quad (8)$$

where SE is the flat disk-shaped structuring element with the radius of 15 pixels. The last process is the noise reduction in BM image by omitting the object that has an area less than 1000 pixels. To ensure that no information is lost related to the cytoplasm, the experiment of selecting the appropriate threshold value in eliminating noise is done through observations of an anatomical pathologist. The final result of this part is shown in Fig. 6.

2) *Segmentation of Overlapping Cytoplasm:* The segmentation of overlapping cytoplasm is important because accurate segmentation of cytoplasm can prevent the loss of important information for the subsequent feature extraction process. Several works had been proposed to separate the overlapping cytoplasm utilize color intensity [16] because the overlapping cytoplasm has lower intensities than the non-overlapping regions. Figure 7(a) shows overlapping cytoplasm region. However, there are overlapping cytoplasm that do not have different intensities to the non-overlapping regions, as shown in Fig. 7(b). This shows that the segmentation method of cytoplasm by only utilizing the color intensity can be false because there is no color intensity difference in the cytoplasm regions.

The method for segmenting the overlapping cytoplasm is proposed as indicated in Algorithm 2.

First, we perform a labeling process of the output image in the cytoplasm boundaries detection stage. Furthermore, from the cytoplasm which has been labeled, we look for the number of nuclei contained

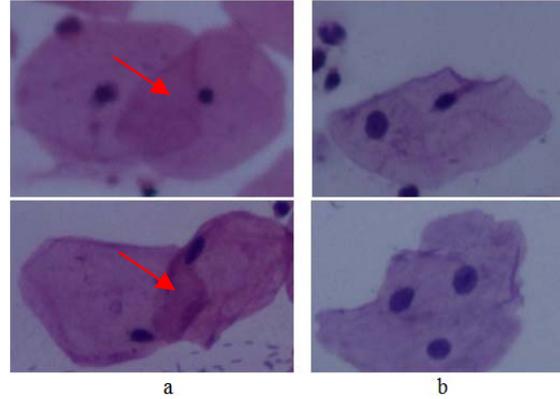


Fig. 7. (a) The image of overlapping cytoplasm that have different color intensity on the overlapping cytoplasm region (indicated by the arrow), (b) The image of overlapping cytoplasm that not have different color intensity on the overlapping cytoplasm region.

Algorithm 2: Segmenting the Overlapping Cytoplasm.

Input : Resulted image with segmentation of overlapping cytoplasm

Output: Pap Smear images with the boundaries of the cytoplasm

Label connected components of Image BM (Cytoplasm Boundaries Detection);

forall the number of labels do

Count the number of nuclei on each label;

if there is more than one nuclei in the cytoplasm region then

Label on the region segmentation of nuclei;

forall the pixels in the cytoplasm region do

Calculate the distance between each pixel in the cytoplasm region with each pixel of nuclei centroid;

Change the value of each pixel on the cytoplasm region with nearby nuclei label values;

end

forall the as much as the number of nuclei in the cytoplasm region do

Doing looping morphological dilation to each region segmentation of nuclei to have areas outside the cytoplasm area;

Combine the results area on stage 1 by area produced on stage 2;

end

end

end

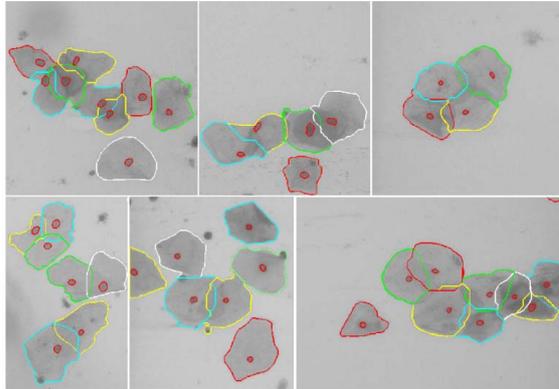


Fig. 8. Representation of segmentation results using proposed method.

in that area. If there is more than one nuclei in a region of the cytoplasm, then the process of calculating the distance between each of the pixel coordinate on the cytoplasmic region and the pixel coordinates on the area centroid nuclei is performed to determine the proximity of each pixel with the nearest centroid. Each pixel in the cytoplasmic region further is classified based on proximity to pixel coordinates of centroid nuclei. The resulting image is the cytoplasmic region that has been classified in accordance with the pixel value (called the image D_c). Furthermore, the process of morphological dilation towards segmentation of nuclei with structuring elements of a matrix shaped nuclei itself is done repeatedly until the area of morphological dilation process results exceeds the area segmentation cytoplasm (called M_d image). Finally, the merger process is performed between the image of D_c and the image of M_d using the 'OR' operator. Figure 8 shows the estimation of segmentation results using the proposed method. The result indicates that the proposed method can segment the overlapping cytoplasm.

III. RESULTS AND DISCUSSION

A. Data

Images used in this study are obtained from the collection of 20 images of pap smear photographs in CITO Laboratory using a NIKON D100 microscope with the two types of preparation: 16 LBP slide and 4 Conventional slide. The number of cells used as the test set is 142 cells consisting of 63 superficial cells, 41 intermediate cells, 29 parabasal cell, and 9 endocervix cell. The images are saved in JPEG format with dimensions of 1280×960 pixels.

TABLE I
THE EXECUTION TIME AT EACH STEP.

Step	Time (s) (mean \pm std)
A. Nuclei Segmentation	2.10 ± 1.16
B. Cytoplasm Segmentation	9.46 ± 2.79

TABLE II
THE RESULTS OF THE t -TEST FOR THE NUCLEI SEGMENTATION.

	Nuclei Area	
	Manual Method	The proposed method
N	142	142
Mean	532.80	532.79
SD	667.13	667.01
Standard error of difference		0.120
t -statistic		0.117
p -Value		0.907
df		141

B. Execution Time

The proposed method requires the execution time of 12.37 ± 3.95 s. Table I shows the time execution of two main steps of the proposed method, namely nuclei segmentation and cytoplasm segmentation. The execution time is presented to test the efficiency of the proposed method. Those methods are built using MATLAB software, 2.66 GHz Pentium of processor and 4 GB of RAM.

C. The t -test

The t -test is used to identify and measure the difference between cytoplasm area generated by proposed method and area generated by manual tracing of experts. The test is applied in both of nuclei area and cytoplasm area. The test used in this study is the paired t -test. It compares two variables of data directly. Those are nuclei area and cytoplasm area. The value of the significance level α is 5%. In other words, the accuracy level of the test results is 95%.

Table II illustrates the t -test result for nuclei area. From Table II, it can be seen that the p -value from the t -test is 0.9070, where that value is greater than $\alpha = 0.05$. It can be concluded that there is no significant difference between the manual tracing by experts and the proposed method.

Table III illustrates the t -test result for cytoplasm area. From Table III it can be seen that the p -value from the t -test is 0.1360, where that value is greater than $\alpha = 0.05$. It can be concluded that there is no

TABLE III
THE RESULTS OF THE t -TEST FOR THE CYTOPLASM SEGMENTATION.

	Nuclei Area	
	Manual Method	The proposed method
N	142	142
Mean	12015.53	12016.31
SD	5971.92	5971.61
Standard error of difference		0.521
t -statistic		1.499
p -Value		0.136
df		141

significant difference between the manual tracing by experts and the proposed method.

D. Hausdorff Distance Test

Hausdorff distance is one of the measurements to compare the distance between two points. Hausdorff distance is used for measuring the similarity level between segmentation image obtained by the proposed method and segmented image traced manually by experts. The Hausdorff distance is defined by [17]:

$$H(A, B) = (h(A, B), h(B, A)) \quad (9)$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \quad (10)$$

The variable A is a set of points (x, y) , and B is a set of points (x, y) of the segmented image resulted by the proposed method. The average value of Hausdorff distance generated in nuclei area is 2.897 ± 3.632 (mean \pm std). Whereas, the average value of Hausdorff distance generated in cytoplasm area is 8.278 ± 6.239 (mean \pm std). Equations (9) and (10) indicate that the smaller the Hausdorff distance is, the higher the similarity degree of two objects. In comparison, Hausdorff distance in cytoplasm area generated by Ref. [16] is smaller than the proposed method, that is 7.69 ± 1.89 . However, the method proposed by Ref. [16] can separate only two overlapping cells, whereas our proposed method can separate more than two overlapping cells. Moreover, in the study done by Ref. [16], there are parameters that must be set manually depending on the dataset used, whereas our proposed method runs automatically without setting specific parameters.

E. Comparison to Another Method

To know its reliability, the proposed segmentation method is compared to the two commonly used

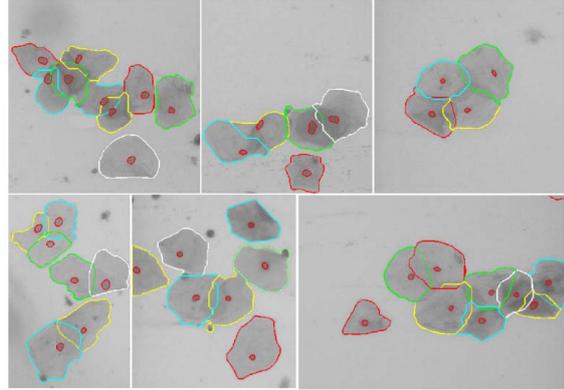


Fig. 9. (a) Cell segmentation using Voronoi Diagrams, (b) Cell segmentation using Watershed Transformation, (c) Cell segmentation using proposed method.

methods: Watershed Transformation [18] and Voronoi Diagram[19]. Figure 9 shows the comparison results of cell segmentation using Voronoi Diagrams, Watershed Transformation, and proposed method. The Voronoi Diagrams can identify cytoplasm region very fast but it is less precise as shown in Fig. 9(a). In this case, Voronoi Diagrams cannot identify the overlapping area. Moreover the wrong identifications as indicated by arrows in Fig. 9(a) are occurred where one cytoplasm is identified as two cytoplasm. This can be occurred because Voronoi diagram use linear approach when dividing space. Meanwhile, in this case, Watershed Transformation can not only segment the cytoplasm precisely, as shown in Fig. 9(b), but there are also variables that must be set manually. Meanwhile, Fig. 9(c) shows the proposed method can segment both the cytoplasm and overlapping cytoplasm precisely.

The proposed method has the advantages to other studies Refs. [16, 20]. These previous studies can only be applied in images that have only two overlapping cells.

IV. CONCLUSION

This study proposes a method based on distance metric and morphological operation to perform an overlapping cytoplasm segmentation. The result shows that the proposed method can segment the overlapping cytoplasm and get cytoplasm boundaries automatically and precisely. In addition, the proposed method is able to separate the boundaries of the cytoplasm without depending on the intensity of color to the cytoplasm. This is considered important because the intensity of colors in the cytoplasm that overlap often have a homogeneous intensity, or cannot be identified the boundaries by naked eye. The results of this study

can be used for other research to improve the accuracy of detection of nuclei in the image of the Pap Smear.

REFERENCES

- [1] WHO. (2012) Cervical cancer estimated incidence, mortality and prevalence worldwide in 2012. Accessed: 26-Jan-2016. [Online]. Available: <http://globocan.iarc.fr/old/FactSheets/cancers/cervix-new.asp>
- [2] 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*, vol. 15, no. 2, pp. 233–241, 2011.
- [3] I. Muhimmah, R. Kurniawan, and Indrayanti, "Automatic epithelial cells detection of pap smears images using fuzzy c-means clustering," in *4th International Conference on Bioinformatics and Biomedical Technology*, 2012, pp. 122–127.
- [4] A. Garrido and N. P. De La Blanca, "Applying deformable templates for cell image segmentation," *Pattern Recognition*, vol. 33, no. 5, pp. 821–832, 2000.
- [5] E. Bak, K. Najarian, and J. P. Brockway, "Efficient segmentation framework of cell images in noise environments," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1. IEEE, 2004, pp. 1802–1805.
- [6] I. Muhimmah, R. Kurniawan *et al.*, "Automated cervical cell nuclei segmentation using morphological operation and watershed transformation," in *Computational Intelligence and Cybernetics (CyberneticsCom), 2012 IEEE International Conference on*. IEEE, 2012, pp. 163–167.
- [7] S.-F. Yang-Mao, Y.-K. Chan, and Y.-P. Chu, "Edge enhancement nucleus and cytoplasm contour detector of cervical smear images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 2, pp. 353–366, 2008.
- [8] K. Li, Z. Lu, W. Liu, and J. Yin, "Cytoplasm and nucleus segmentation in cervical smear images using radiating gvf snake," *Pattern Recognition*, vol. 45, no. 4, pp. 1255–1264, 2012.
- [9] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," in *Proceedings of Nature inspired Smart Information Systems (NiSIS)*, 2005, pp. 1–9.
- [10] N. Mustafa, N. A. M. Isa, and M. Y. Mashor, "Automated multicells segmentation of thinprep image using modified seed based region growing algorithm (special issue on biosensors: Data acquisition, processing and control)," *Biomedical fuzzy and human sciences: the official journal of the Biomedical Fuzzy Systems Association*, vol. 14, no. 2, pp. 41–47, 2009.
- [11] R. S. Hoda and S. A. Hoda, *Fundamentals of Pap test cytology*. Springer Science & Business Media, 2007.
- [12] I. Muhimmah, R. Kurniawan *et al.*, "Analysis of features to distinguish epithelial cells and inflammatory cells in pap smear images," in *The 6th International Conference on Biomedical Engineering and Informatics*, 2013, pp. 519–523.
- [13] T. Ridler, S. Calvard *et al.*, "Picture thresholding using an iterative selection method," *IEEE trans syst Man Cybern*, vol. 8, no. 8, pp. 630–632, 1978.
- [14] R. Kurniawan, "Modified watershed algorithm based on distance-metric criterion for nuclei clustered separation in pap smear images," *Jurnal Teknoin*, vol. 19, no. 1, 2016.
- [15] R. S. Gonzalez and P. Wintz, *Digital Image Processing*. Addison-Wesley Publishing Company, 1977.
- [16] M. E. Plissiti, M. Vrigkas, and C. Nikou, "Segmentation of cell clusters in pap smear images using intensity variation between superpixels," in *Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on*. IEEE, 2015, pp. 184–187.
- [17] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–863, 1993.
- [18] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2405–2414, 2006.
- [19] D. M. Ushizima, A. G. Bianchi, and C. M. Carneiro, "Segmentation of subcellular compartments combining superpixel representation with voronoi diagrams," *Overlapping Cervical Cytology Image Segmentation Challenge-IEEE ISBI*, pp. 1–2, 2014.
- [20] S. N. A. M. Kanafiah, Y. Jusman, N. A. M. Isa, and Z. Mohamed, "Radial-based cell formation algorithm for separation of overlapping cells in medical microscopic images," *Procedia Computer Science*, vol. 59, pp. 123–132, 2015.