MODIFIED WATERSHED ALGORITHM BASED ON DISTANCE- METRIC CRITERION FOR NUCLEI CLUSTERED SEPARATION IN PAP SMEAR IMAGES

Rahadian Kurniawan

Department of Informatics, Universitas Islam Indonesia, Yogyakarta
rahadiankurniawan@ymail.com

ABSTRACT

An improved watershed algorithm based on distance-metric criteria to separate overlapping cells proposed in this study. In this study, we present a novel watershed-based method for separation of overlapping cervical cell in Pap smear images. The results will be compared with other method namely distance transform, and color morphology gradient. The result of proposed method shows not only produces accurate separating cells cluster, but also keeps the shape of cells in good form.

Keywords: Watershed, Cervical Cancer, Pap Smear.

1. INTRODUCTION

The high degree of cell overlap, pose a challenge to any method to address the complexity of cervical cell image (Pap smear) and to achieve a correct segmentation. The automation separation process with the computer assisted of these overlapping cells will be very helpful to farther cell analysis. For example, the separation of overlapping cells is an early stage to make the process of counting the number of cells. If the process fails, some cell separation will only be counted as one cell. Therefore, the process of separation of overlapping cell is very helpful for the further cell analysis. The separation process is not only able to separate cells properly, but should be able to maintain cell form in good shape. This process is an important process in the cell segmentation stage, and can contribute to distinguish normal and abnormal cells in Pap smear images.

In order to solve the problem of over-segmentation, people have proposed several improved methods for watershed, among other: distance transformation [1], color morphological gradient [2]. Similar studies are conducted to counting the number of nuclei in Pap smear images [11] and segment nuclei in Pap smear images have been proposed, such as: Watershed-based [3], and morphological operations and watershed transformation [4]. In addition, similar studies are conducted to separation of clustered nuclei, include: H - minima Transform - Based Marker Extraction and Contour Parameterization [5], another methods concerns morphological watersheds for the separation of the cytoplasm and the nucleus of each cell [6], [7]. There are many methods proposed in previous studies, but the method of cell separation is most commonly used watershed algorithms. This method is most commonly used because of the computing process is fast and capable of separating cells precisely. However, the traditional watershed algorithm is very sensitive to noise that can cause errors in the separation process.

This study proposes an algorithm to enhance the ability of the watershed algorithm to separate overlapping cells using distance criteria. The results will be Compared with [1] and [2]. In the rest of the paper, Section II presents several theories related to the proposed method. Section III presents the proposed method. In section IV, presents the experiments and evaluation result and also discusses the proposed method, Section V provides the conclusion of the proposed method.

2. THEORY

2.1. Watershed Algorithm

Watershed transformation is a morphological based tool for image segmentation. The watershed transformation is originally proposed by Digabel and Lantuejoul [8]. The way to explain watershed segmentation is the Immersion Approach (see fig 1): Catchment Basins (CB) will be filled up with water starting at each local minimum. If the water of different catchment basins is likely to meeting due to further immersion, a dam is built to avoid the merging of catchment basins. This flooding process will eventually reach the highest level where the water can reach in the landscape. As a result, watershed lines can separate individual catchment basins in the landscape.
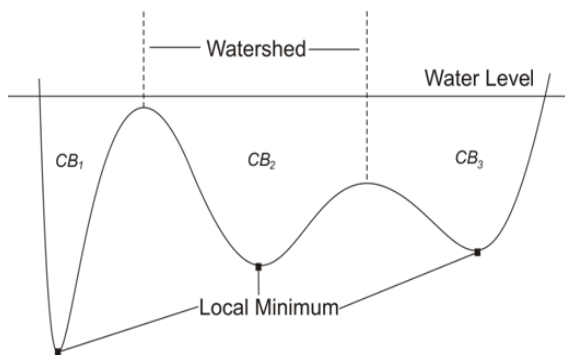

Fig. 1.   Illustration of Immersion Approach.

General watershed process can be divided into two parts: the first is sorting and the second is flooding. An efficient algorithm to implement this approach proposed by Luc and Soille [9]. Generally, the algorithm includes two steps:

1. *Sorting.* The gray value of each pixel is sorted in the image.

2. *Flooding.* The flooding begin from the minimum of gray level h. X_h is made of the points of image which belong to the minima of lowest altitude. Now, we consider the gray level $h+1$. Y_{h+1} being one of the connected component of the pixel of gray level $\leq h + 1$, there are three possible relation of inclusion between Y_{h+1} and $Y_{h+1} \cap X_h$:

a. $Y_{h+1} \cap X_h = \emptyset$; : In this case, Y_{h+1} is obviously a new minimum of at gray level $h+1$;

b. $Y_{h+1} \cap X_h \neq \emptyset$; and it is connected. : in this case, Y_{h+1} corresponds exactly to the pixels belonging to the catchment basin associated with the minimum of X_h

c. $Y_{h+1} \cap X_h \neq \emptyset$; and it is not connected. : in this case, we can infer that Y_{h+1} must contain different minima at gray level h.
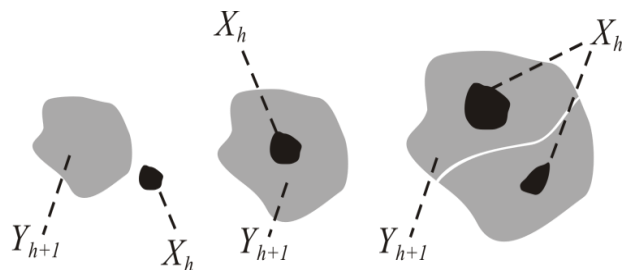

Fig. 2.   The Three Possible Inclusion Relations Between X_h And Y_{h+1}

2.2. Distance Transform (DT)

The approach based on the distance transformation is the type most commonly used to improve the performance of watershed algorithm. This method changes the location information of pixels in an image into gray information, so images are suitable for watershed algorithm. Because this method is largely dependent on location information, are not suited for complex situations. For example, when the number of cells that are used a lot and many of them are abnormal cells, over-segmentation problem still occurs frequently and sometimes this method fails to work properly. There are many different ways to define the distance between (x_1, y_1) and (x_2, y_2) in a digital image. Several commonly used DT functions for image processing are shown in table 1.

Table 1 Distance Transform Formula

Method	Formula	
Chessboard	$\max\left(\lvert x_1 - x_2\rvert, \lvert y_1 - y_2\rvert\right)$	(1)
City block	$\lvert x_1 - x_2\rvert + \lvert y_1 - y_2\rvert$	(2)
Euclidean	$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$	(3)
Quasi-Euclidean	$\lvert x_1 - x_2\rvert + \left(\sqrt{2}-1\right)\lvert y_1 - y_2\rvert, \lvert x_1 - x_2\rvert > \lvert y_1 - y_2\rvert \left(\sqrt{2}-1\right)\lvert x_1 - x_2\rvert + \lvert y_1 - y_2\rvert, otherwise$	(4)

2.3. Color Morphological Gradient Image

The color morphological gradient is an extension of the morphological gradient of a grayscale image in color space. The classical definition of the morphological gradient of a grayscale image is terms of grayscale dilation $\delta_g(f)$ and grayscale erosion $\varepsilon_g(f)$ as:

$$\nabla(f) = \delta_g(f) - \varepsilon_g(f), \quad\quad (5)$$

For the expression of Morphological gradient can be made in other ways that is the maximum absolute intensity difference between two pixels in the area of the structuring element *G*, as follows:

$$\nabla(f) = \max_{x \in G}\{f(x)\} - \min_{x \in G}\{f(x)\} \quad (6)$$

$$= \max(\lvert f(x) - f(y)\rvert) \forall x, y \in G$$

In the color image, the pixel is represented as a three-dimensional vector with elements of the intensity of the three component colors and color gradients morphology (M_{Grad}) can be expressed as:

$$M_{Grad} = \max_{i,j \in G}\left\{\left\lVert x_i - x_j\right\rVert_p\right\} \quad (7)$$

where x_i, x_j are pixels in the structuring element *G*. In this study the second norm ($p = 2$) was calculated and a 3×3 flat structuring element was used.

3. METHOD

Nuclei region has an intensity which is not homogeneous (see Figure 3 (a)). Figure 3 (b) is the intensity map of the image 3 (a), where the red color represents high intensity, and blue colors represent low intensity. The arrows in Figure 3 (c) shows the nuclei have a lot of intensity valley (region minima). In Figure 4, seen that in the image segmentation results nuclei contained several region minima marked with a dot (o). From the Figure, for nuclei with labels 7 and 5 have more than one region minima.

While, on the other nuclei (another label) has only one region minima.
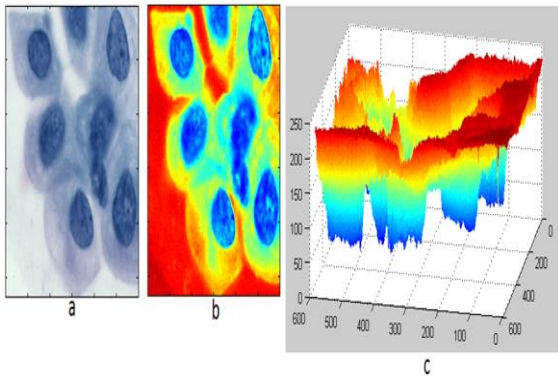

Fig. 3.   (a) Initial Cell Image. (b) Mapping of The Intensity Values In The Color Space, Where High-Intensity Values are Represented by Red And Small-Intensity Values Are Represented by Blue. (c) Mapping of The Initial Image In 3-D Space. The Arrows Indicate The Nuclei Have a Lot of Regional Minima.
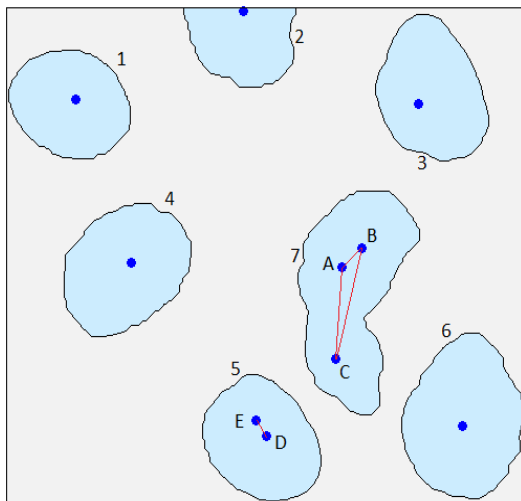

Fig. 4.   Segmentation Results Nuclei Contained Several Region Minima Marked With a Dot (o).

The expected results of the proposed method is a nuclei that have only one region minima, which is used as the point of the basin for the cell separation process using watershed algorithm. If the nuclei in a region there are more than one regional minima, it will look for the distance between the centroids that represent the area such minima region. If the distance between the two centroids is less than the value of T, (in this study the value of $T = 35$) then it will be compared to the intensity of color in both centroids. Centroids with a lower intensity of the color will last, and centroid with high color intensity will be eliminated. The following algorithm describes the process that occurs.

1.   Input images *s;*
2. $g \leftarrow$ labels on the image s;
3. $NL \leftarrow$ number label of the image g (step 2);
4. $Cb \leftarrow$ perform distance transform with city-block method from *g* (step 2);
5. $Rm \leftarrow$ region minima from Cb (step 4);
6. $F \leftarrow XOR(Rm, s);$
7. $Im(x,y) \leftarrow \begin{cases} 0, & F(x,y) = 0 \\ g(x,y), & F(x,y) = 1, \end{cases}$ x is the abscissa and y are ordinate of pixels in the image;
8. $i=1$;
9.   Repeat;
10. $Nc \leftarrow$ number of centroids from *Im* (step 7), where $\exists\, I(x,y) = i$;
12.   if Nc > 1;
13. $temp = \emptyset$;
14.   repeat;
15. $\forall p = (x,y) \in R_c$;
16.   if exist $q = \{(x_q, y_q) \mid D(p,q) \leq T\}$;
17. $temp = \{temp, p, q / \max \{I(p), I(q)\}\}$;
18.   *until p evaluated;*
19. $Utemp = unique(temp)$;
20. $RcNew = Rc \,\Delta\, Utemp$;
21.   *end;*
22.   *i++;*
23.   until i=NL;
24.   Output *RcNew;*

The image of *s* is a binary image segmentation results (in this study can be seen in [4]). The first step is the labeling process [12] to the image of *s*. Then, we find the number of labels from the image *g*. From the image segmentation results, we find the distance of each pixel value *J* with city-block method. Thus, we find region minima from *Cb*. The output image of the region minima is a binary image that would be combined with image segmentation result s using the XOR operator. Furthermore, if the output image in step 6 (*F*) has a value in pixels = 1, then converted to the value at the same pixel

in the image of g. However, if the pixel value = 0, then the value is still filled with 0.

Rc is the set of centroids obtained from the image of *F* (step 6), *D* is the Euclidean distance between the two centroids, *T* is the threshold value of the minimum distance, *I* (*p*) is the image intensity at the centroids *p*, and *temp* is a temporary storage variable. *Temp* variable will take a value between two high color intensity. This means a lower intensity persisted, whereas higher intensities included in the *temp*. Centroids reduction will be minimal if *T* is set low, but if the value of *T* at high set can allow centroids that are appropriate in the nuclei become lost. The algorithm is repeated until entire *p* evaluated, all centroids have been evaluated by a combination of the *q*. The total number of combinations between 2 centroids obtained by:

$$_nC_k \equiv \binom{n}{k} \equiv \frac{n!}{k!(n-k)!} \quad \ldots\ldots\ldots\ldots(8)$$

where *n* is the number of centroids, while *k* = 2 since it only involves a combination between 2 centroids. *Utemp* is a variable that contains the value of the set *temp* not redundant or more than one. *RcNew* is the symmetry difference (Δ) between *Rc* and *Utemp*. *RcNew* values are the coordinates of the new centroids.

Algorithm above will generate a point on each nuclei (see Figure 5 (b)). The point will be used as a marker of regional minima. Furthermore, we can perform a watershed transformation to separate the nuclei segmentation. Figure 5 shows the results of the algorithm. As a note, the result of nuclei segmentation in this study was obtained from previous studies [4].
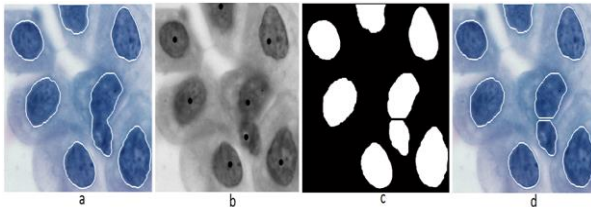

Fig. 5. (a) Initial Image With Nuclei Boundary, (b) Superimposing The Regional Minima on

Grayscale Image, (c) Result From Watershed Algorithm Based on Proposed Method, (d) Initial Image With Nuclei Boundary From Proposed Method.

Figure 6 shows the results of the proposed method to another dataset of Pap smear image used in this study. From Figure 6 shows the advantages of using the proposed method which is not only able to separate the cells with precision, but also maintaining good form of segmentation of nuclei.
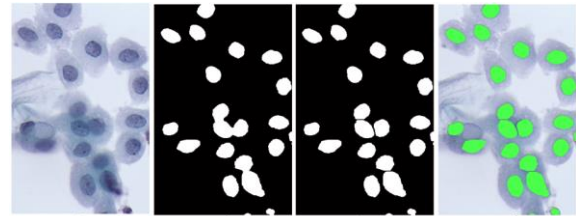

Fig. 6. (a) Initial Image, (b) Initial Image With Nuclei Segmentation Boundary, (c) Separation of Clustered Nuclei With The Proposed Method, (c) Resulted Proposed Method.

4. RESULT AND DISCUSSION

Total data used in this study are 9 image of the Pap smear image that is divided into training set and test set. Training set is used to find the value of *T* which is ideal as a minimum threshold value of distance in the proposed method, while the test set is used to test the results of the proposed method. The test set is composed on 6 pap smears images of NCI Bethesda System [10] with different variations in preparation type. These images contain 140 nuclei and and have been confirmed by our pathologist. The size of the images is 2000 × 1600 and they were stored in JPEG format. In this paper, we compare our proposed method with other method namely, Distance Transform [1] and Color Morphological Gradient [2]. Figure 7 shows the results of a comparison between the proposed method with others methods. The proposed method result shows better nuclei segmentation separation than other method.
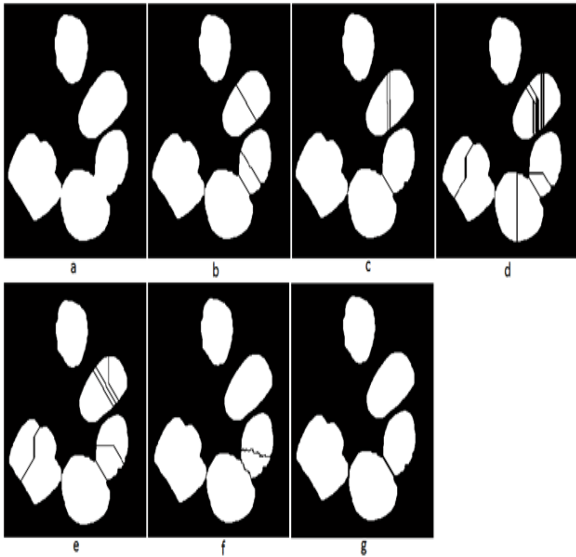
(a) Binary Image Which is Obtained After The Preprocessing, (b) Result from Watershed Algorithm Based on Distance Transformation Using Chessboard Method, (c) Result from Watershed Algorithm Based on Distance Transformation Using Cityblock Method,(d) Result from Watershed Algorithm Based on Distance Transformation Using Euclidean Method,(e) Result from Watershed Algorithm Based on Distance Transformation Using Quasi-Euclidean Method, (f) Result from Watershed Algorithm Based on Color Morphological Gradient Transformation, (g) Result from Watershed Algorithm Based on Proposed Method.

Distance parameter (T value) between U and V vector obtained by the formula $D^{Euc} = \sqrt{(u-v)^T(u-v)}$. Distance threshold value in this study is obtained through rigorous testing to 3 images of the dataset used (training set). Figure 8 shows the graph tests conducted to determine the distance threshold value (T).
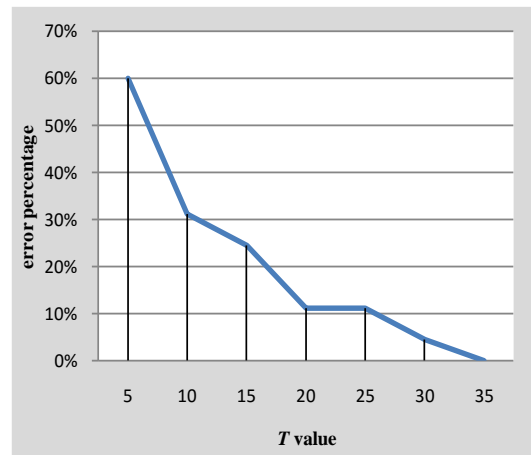

Fig. 8.Percentage Error Based on The Value of T

The T value ($T = 35$) is considered the most ideal value for our dataset. The greater value of T has the possibility of loss of true regional minima. Should be noted that the image is used to measure the value of T are different from the image that is used to test the results of the proposed method. This is accomplished to see the level of robustness of the value of T if another image is used. Images that used in testing the value of T consists of 3 Pap smear images containing 45 cell nuclei with 12 nuclei which are overlapping. Furthermore, as a measure of the computational efficiency of the proposed method, we present in Table 2 comparison of the proposed method and other methods. The proposed method developed in MATLAB using a dual-core 2.80 GHz and 2 GB of RAM.

Table 2 Execution Time

Method	Mean (sec)	Stdev (sec)
City block DT	1.61	0.62
Chessboard DT	1.72	0.66
Euclidean DT	1.55	0.61
Quasi-Euclidean DT	1.75	0.68
Color Morphological Gradient Image	4.88	1.90
Proposed	7.17	3.58

Comparison in Table 2 shows that the Distance Transform is a method that is relatively fast compared to the proposed method, although not significantly. This is because the computational speed of the proposed method is strongly influenced by the number of labels contained in the image. However, the resulting segmentation results show that the proposed method is much better than the Distance Transform and morphological Color Gradient. The following Table 3 show the comparison results of counting the number of cells manually with separation methods based on cell segmentation algorithm watershed.

It can be seen from the Table 3 that the proposed method showed much better results than other methods. From the six images used, the proposed method is able to perform separation of overlapping cell segmentation well. It is shown that the results of the percentage of error are extremely small. Errors that occur when using the proposed method only happen in Figures 1 and 6. Percentage negative error indicates that there is a failure of cell separation using the proposed method. The negative error means the number of calculation from proposed methods are less than the manual counting. Because the percentage of positive error indicates an error that causes cell truncation errors, so the number of calculation from proposed methods are used more than the number of manual counting.

Table 3. Number of Components (NOC) and Errors Produced by Different Method.

File	NOC by pathologist manual count	NOC by cityblock DT	NOC by chessboard DT	NOC by euclidean DT	NOC by quasi-euclidean DT	NOC by color morphological gradient image	NOC by Proposed Method
1	22	23 Error: +4.5%	24 Error: + 9.1%	38 Error: + 72.7%	32 Error: + 45.5%	25 Error: + 13.6%	21 Error: -4.5%
2	21	32 Error: + 52.4%	25 Error: + 19.0%	62 Error: + 195.2%	54 Error: + 157.1%	25 Error: + 19.0%	21 Error: + 0.0%
3	40	50 Error: + 25.0%	49 Error: + 22.5%	102 Error: + 155.0%	78 Error: + 95.0%	46 Error: + 15.0%	40 Error: + 0.0%
4	21	36 Error: + 71.4%	25 Error: + 19.0%	64 Error: + 204.8%	51 Error: + 142.9%	23 Error: + 9.5%	21 Error: + 0.0%
5	5	7 Error: + 40.0%	7 Error: + 40.0%	13 Error: + 160.0%	10 Error: + 100.0%	6 Error: + 20.0%	5 Error: + 0.0%
6	31	42 Error: + 35.5%	39 Error: + 25.8%	78 Error: + 151.6%	70 Error: + 125.8%	34 Error: + 9.7%	29 Error: -6.5%

The failure of cell separation process by the proposed method is due to the overlapping nuclei regions only have a single region minima. Figure 9 shows the segmentation cells in an overlapping region only has a single minima. The arrows in Figure 9 shows the overlapping nuclei region but only detected as single minima (denotes with a ret dot).
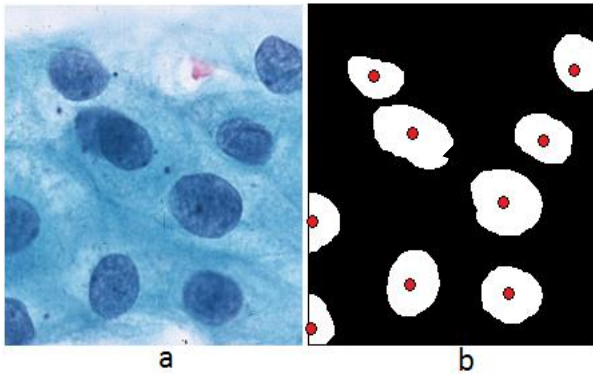
Fig. 7. (a) Initial Image, (b) Segmented Nuclei with Region Minima, Denotes with a Ret Dot (o).

5. CONCLUSION

In this paper we propose an improved watershed algorithm based on distance-metric criterion. As it is verified by the results, the method is suitable for improving the detection and calculation number of cell nuclei in Pap smear images. The experiment shows that the proposed algorithm not only produces an accurate separating cells cluster, but also keeps the shape of cells in good form. On the other hand, the algorithm is implemented automatically that works to the computations efficiently. It must be noted that the distance parameters must be defined to reach acceptable results.

REFERENCES

A. N. Evans, "Morphological gradient operators for colour images", *Proceedings of the IEEE International Conference on Image Processing (ICIP04)*, Vol. 5, pp. 3089-3092, 2004.

C. Jung and C. Kim, "Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2600–2604, Oct. 2010.

Digabel, H., & Lantuejoul, C. (1977). Iterative algorithms. In J.-L. Chermant (Ed.), Actes du Second Symposium Europeen d'Analyse Quantitative des Microstructures en Sciences des Materiaux, Biologie et Medecine, Caen, 4-7 October 1977, pp. 85-99. Stuttgart: Riederer Verlag.

Haralick, Robert M., and Linda G. Shapiro, Computer and Robot Vision, Volume I, Addison-Wesley, 1992, pp. 28-48.

Luc Vincent and Pierre Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, Vol. 13, No. 6, June 1991, pp. 583-598.

M. E. Plissiti, C. Nikou and A. Charchanti, "Watershed-based segmentation of cell nuclei boundaries in Pap smear images", *Proceedings of the 10th IEEE International Conference on Information Technology Applications in Biomedicine (ITAB'10)*, 3-5 November 2010, Corfu, Greece.

Muhimmah, I, Kurniawan, R, Indrayanti, "Automatic Epithelial Cells Detection of Pap smears images using Fuzzy C-Means Clustering", *2012 4th International Conference on Bioinformatics and Biomedical Technology, pp. 122-127*, 2012.

Muhimmah, I., Kurniawan, R., Indrayanti (2012) Automated Cervical Cell Nuclei Segmentation using Morphological Operation and Watershed Transformation. *In IEEE International Conference on Computational Intelligence and Cybernetics* 2012.

NCI Bethesda System Web Atlas, http://nih.techriver.net/.

O. Lezoray and H. Cardot, "Cooperation of color pixel classification schemes and color watershed: A study for microscopic images," IEEE Trans. Image Process., vol. 11, no. 7, pp. 783–789, Jul. 2002.

P. Bamford and B. Lovell, "A water immersion algorithm for cytological image segmentation," in Proc. APRS Image Segmentation Workshop, Sydney, Australia, 1996, pp. 75–79.

The Mathworks online documentation, Matlab Image Processing Toolbox, http://www.mathworks.com/help/images/ref/bwdist.html?searchHighlight=distance+transform