Automated Blasts Segmentation Techniques Based on Clustering Algorithm for Acute Leukemia Blood Samples

Nor Hazlyna Harun¹, Mohd Yusoff Mashor¹ and Rosline Hassan²

¹ Electronic & Biomedical Intelligent Systems (EBItS) Research Group, School of Mechatronic Engineering, Universiti Malaysia Perlis, Perlis, Malaysia
² Hematology Department Universiti Hospital, Universiti Sains Malaysia, Kelantan, Malaysia

¹hazlyna_harun@yahoo.com, ¹yusoff@unimap.edu.my, ²roslin@usm.edu.my

ABSTRACT

Image segmentation is an important process for most medical image analysis tasks. One of the familiar segmentation technique is using clustering algorithm. At present, clustering algorithm has been used in many fields including machine learning, data mining, pattern recognition, image processing, and bioinformatics. These methods partition or group the objects based on similarities and differences. This study aims to assist hematologist or technologist by automating the visual assessment to identify blasts in acute leukemia blood samples. In this study, an automated image segmentation technique using a combination between saturation component based on HSI color space and clustering algorithm which are Moving k-Means and Fuzzy c-means are proposed. Then, 7x7 pixels median filter and region growing algorithms was applied to remove the unwanted noises. The comparison performance of the proposed technique was investigated. The experimental results yield a promising result for the combination between saturation component with Moving k-means clustering algorithm, 7x7 pixels median filter and region growing algorithms.

Keywords: Acute leukemia blood images, clustering, Moving k-means, Fuzzy c-means, saturation formula.

1. Introduction

The term leukemia refers to a group of cancers of the blood cells. In leukemia, white blood cells become abnormal, divide and grow in an uncontrolled way. These cells are start in the soft, inside part of the bones called the bone marrow. After it starts, leukemia cells often shifts quickly into the blood where it can reach to other parts of the body such as the lymph nodes, spleen, liver, central nervous system (brain and spinal cord), and other organs (Lim, 2002).

According to the National Cancer Institute report in 2009, for 17 SEER (Surveillance Epidemiology and End Results) geographic areas, it is estimated 5,330 men and women (3,150 men
and 2,180 women) will be diagnosed with acute lymphocytic leukemia (ALL) in 2010. The statistic also estimated 12,330 men and women (6,590 men and 5,740 women) will be diagnosed with acute myeloid leukemia (AML) in 2010. In addition, 8,950 men and women will die due to acute myeloid leukemia while 1,420 men and women will die due to acute lymphocytic leukemia in 2010. Leukemia is a complex disease, and there are two main types of leukemia: acute leukemia and chronic leukemia. The acute leukemia is divided into two categories, depending upon their cell of origin. Leukemia evolving from the myeloid/granulocyte cell line is called acute myelogenous leukemia (AML). Lymphocytic precursors give rise to acute lymphocytic leukemia (ALL) (Lim., 2002). The word acute means that the cancers grow rapidly, and if not treated might be serious in a few months. The original classification scheme proposed by the French-American-British (FAB) Cooperative Group which is differentiated based on morphology, including cell size, prominence of nucleoli, and the amount and appearance of cytoplasm (Mittal & Meehan, 2001). According to French-American-British (FAB) classification also, the description of cells is small and uniform for ALL-L1. Meanwhile, cells of AML-M1 are large and regular (Mittal & Meehan, 2001). In laboratories, hematologists and technologists investigate human blood by microscopic investigation. Since this is done by humans, it is not fast and it does not present a standardized accuracy caused by the operator’s capabilities and tiredness (Piuri & Scotti, 2004). The reproducibility of the results is sometimes misinterpreted and lacking. In order to assist medical persons to improve the reliability of the results, the automated image processing system was introduced and offers useful tools for the medical field, especially for investigating acute leukemia diseases.

Due to difficult nature of the blood cell images and difference in slide preparation many effort have to be done to meet the current clinical demands. The success of an automated image processing system of blood cells depends on the proper segmentation of images. Currently, there are many established segmentation techniques used in medical image such as thresholding (Cseke, 1992; Liao & Deng, 2002), cell modeling (Liao & Deng, 2002; Jiahng et. al, 2003), watershed clustering (Jiahng et. al, 2003; Venkateswaran & Ramana Rao, 2007), mathematical morphology (Anoraganingrum, 1999) and clustering algorithm (Piuri & Scotti, 2004). Having good segmentations will give an advantage to clinicians as it can provide essential information for early disease detection (Ng, 2006). Clustering algorithm is an unsupervised task that can divide the given set of data into several non-overlapping of homogenous groups (Venkateswaran & Ramana Rao, 2007). Besides that, the clustering technique normally represent image into cluster space (Rezaee et al., 1995). In medical image segmentation task, the used of clustering algorithm is easier and more effective, since the number of clusters for the structure of interest is usually known from its anatomical information (Sulaiman & Isa, 2010). Clustering algorithm has been applied as a digital image segmentation method in different fields including computer, engineering, and mathematics. Furthermore, the clustering algorithms are widely used and have been proved to produce good performance in medical image segmentation such as k-means, Fuzzy c-means and Moving k-means. These include the works presented in (Chen et. al, 1998; Pham et.al, 1997; Mat-Isa et.al, 2003). In order to automatically segment the blasts in acute leukemia blood samples, this study focus on two types of clustering algorithms ability, which are Moving k-means and Fuzzy c-means. In addition, to ease the automatic segmentation process, the clustering algorithm has been combined with saturation formula base on Hue Saturation Intensity (HSI) color space. Then, 7x7 pixels median filter and seed region growing algorithms have been applied to remove unwanted noises. After that, the proposed auto segmentation techniques performance has been evaluated using pixels subtraction technique and quantitative assessment base on manual segmentation image as ground truth image.
2. Methodology

The proposed technique consists of saturation formula base on HSI color space, clustering algorithm and 7x7 pixels median filter. The methodology involves the following steps:

2.1 Image Acquisition

The acute leukemia blood slide images were analyzed using Leica microscope at 40 magnifications. Then, Infinity 2 camera was used to capture images and saved into (.bitmap) format with 800x600 resolution. The acute leukemia blood slide samples were provided by Hospital Universiti Sains Malaysia (HUSM) Kubang Kerian Kelantan.

2.2 Threshold Selection Method

The threshold selection method consists of two steps. The proposed selection method will utilize saturation formula based on HSI color space. Then, the resulted image will be used as the input for clustering algorithm. The steps are:

2.2.1 Saturation base on HSI color space

Generally, the HSI color space consists of three components which are Hue, Saturation and Intensity. The saturation component measures the degree of white light added to pure color. Based on observation of blood cells image, the acute leukemia cells (blast) are the most highlighted and clearly be seen in saturation component image. Meanwhile, the red blood cells and other particles became less saturated in the saturation component image. Therefore, the saturation component was chosen in order to reduce the computational effort and to ease the clustering process. The saturation formula has been applied to the original acute leukemia blood slide images as in

\[ \text{Sat} = 1 - \frac{3}{R + G + B} \min(R, G, B) \]

(1)

2.2.2 Clustering Technique

In this study, two clustering algorithms namely Moving k-means and Fuzzy c-means have been used for automated segmentation purpose. These two clustering algorithms have been applied to the resultant saturation images and the performance will be compared at the end of this paper.

2.2.2.1 Moving k-means clustering

In 2000, Mashor proposed a modified version of k-means algorithm, which is moving k-means algorithm. Moving k-means algorithm is capable to minimize dead centers, pixel redundancy problems and the effect of trapped in local minima problems (Mashor, 2000). Based on original Moving k-means clustering algorithm, the algorithm of Moving k-means clustering can be implemented as:

1. Initialize the centers and \( \alpha_0 \), set \( \alpha_a = \alpha_b = \alpha_0 \), (where is \( \alpha_0 \) small constant value, \( 0 < \alpha_0 < \frac{1}{3} \) and should be chosen to be inversely proportional to the number of centre).
2. Assign all pixels to the nearest centre and calculate the center position as in
\[ C_j = \frac{1}{n_j} \sum_{i \in C_j} v_i \quad (2) \]

3. Check the fitness of each center using as in
\[ f(C_j) = \sum_{i \in C_j} (|v_i - C_j|)^2 \quad (3) \]

4. Find \( C_s \) and \( C_l \), the cluster that has the smallest and the largest value of \( f(.) \)

5. If \( f(C_s) < \alpha_a f(C_l) \)

5.1. Assign the pixels of \( C_i \) to \( C_s \) if \( v_i < C_s \) and \( i \in C_i \) and leave the rest of the pixels to \( C_l \)

5.2. Recalculate the positions of \( C_s \) and \( C_l \) as in
\[ C_s = \frac{1}{n_s} \sum_{i \in C_s} v_i \quad (4) \]
\[ C_l = \frac{1}{n_l} \sum_{i \in C_l} v_i \]

Note: \( C_s \) will give up its pixels before step (5.1) and, \( n_s \) and \( n_l \) are the number of the new pixels of \( C_s \) and \( C_l \) respectively, after the reassigning process in step (5.1).

6. Update \( \alpha_a \) according to \( \alpha_a = \alpha_a - \frac{\alpha_a}{n_s} \) and repeat step (4) and (5) until \( f(C_s) \geq \alpha_a f(C_l) \).

7. Reassign all pixels to the nearest centre and recalculate the centre positions as in (2).

8. Update \( \alpha_a \) and \( \alpha_b \) according to \( \alpha_a = \alpha_b \) and \( \alpha_a = \alpha_b - \frac{\alpha_b}{n_l} \) respectively, and repeat step (3) to (7) until \( f(C_s) \geq \alpha_b f(C_l) \).

9. Sort the centers in ascending order where \( C_1 < C_2 < \ldots < C_n \).

2.2.2.2 Fuzzy c-means clustering

Fuzzy c-means is developed by Dunn in 1973 and improved by Bezdek in 1981. Fuzzy c-means clustering minimizes the following function as in
\[ J = \sum_{i=1}^{N} \sum_{j=1}^{N} (M_{ij})^m ||v_i - C_j||^2 \quad (5) \]

where, \( M_{ij} \) is the partition matrix, which represents the degree of membership between each data sample and all centers and \( m \) is any real number greater than one. \( N \) is the number of cluster, \( n_s \) is the number of data, \( v_i \) is \( i \)-th sample of data and \( C_j \) is \( j \)-th center of cluster. The value of \( m \) will controls the fuzziness of the membership function. Based on Bezdek J.C. in 1981 for \( m = 2 \), the Fuzzy c-means algorithm is composed of the following steps:

1. Initializes the centers.
2. Calculate \( M_{ij} \) as in (6) and (7).
\[ M_{ij} = \frac{1}{\sum_{k=1}^{n} \left( \frac{d_{ij}}{d_{kj}} \right)^2} \text{ if } d_{ij} > 0, \forall i,j \]  

(6)

where \( d_{ij} = \| v_i - c_j \|^2 \).

\[ M_{ij} = 1 \]

\[ M_{ij} = 0 \text{ for } i \neq 1 \]

(7)

3. Calculate the new position of the centers as in

\[ C_j = \frac{\sum_{i=1}^{N} M_{ij}^m v_i}{\sum_{i=1}^{N} M_{ij}^m} \]  

(8)

4. Repeat steps (2) and (3) until the centers no longer move.

2.3 Removing unwanted noise

After completing the clustering process, the outputs of each type of clustering techniques were filtered by using 7x7 pixels median filter and seed region growing algorithms in order to improve the resulted images. This section is further divided into two steps are: 7x7 pixels median filter and seed region growing algorithms.

2.3.1 7 x 7 Pixels Median Filter Algorithm

The 7 x 7 pixels median filter has been applied to the segmented images in order to remove unwanted noise such as salt and pepper noises. However, it can be noted that 7 x 7 pixels median filter could not totally removed the unwanted noise that are bigger in size.

2.3.2 Seed Region Growing Algorithm

After the resulted images were filtered, there are still the unwanted noises which area bigger in size that is unable to be filtered by using 7 x 7 pixels median filter. Therefore, seed region growing algorithm is applied to the image resulted from 7 x 7 pixels median filter. This process yields better and more acute leukemia blood cells image visualization.

3. Results And Discussions

The proposed method described above are tested on images for acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL), which were taken from seven slides of acute leukemia blood samples. Figures 1 (a) - 4 (a) show the original captured acute leukemia images with resolution of 800x600. Meanwhile, Figures 1 (b) - 4 (b) show the images after applying the saturation formula. The results obtained in Figures 1 (b) - 4 (b) show that the proposed method by beginning with applying the saturation formula to AML and ALL original images to ease the clustering process. Figures 1 (d) - 4
(d) and Figures 1 (e) - 4 (e) represent the output images using Fuzzy c-means and Moving k-means algorithm.

From the results shown in Figures 1 (d) – 2 (d) and Figures 1 (e) -2 (e) for AML images, as well as Figures 3 (d) - 4 (d) and Figures 3 (e) - 4 (e) ALL images, it can be noticed that the two types of clustering algorithm can cluster the region of interest for blast in the acute leukemia blood images into three regions which are background, nucleus and cytoplasm. However, it still can be seen that the presence of unwanted noises from the Figures 1 (d) - 4 (d) and Figures 1 (e) - 4 (e).

Figures 1 (f) - 4 (f) and Figures 1 (g) - 4 (g) show the output images after filtering using 7x7 pixels median filter. At this time, most of the salt and pepper noise existed previously have already been eliminated. Anyhow, there are still unwanted noises which are bigger in size that are unable to be filtered using 7x7 pixels median filter. Therefore, in order to give better visualization results, seed region growing algorithm has been further applied. After the region growing algorithm has been applied, from the Figures 1 (h) - 4 (h) and Figures 1 (i) - 4 (i), the output images appear to be clearer and cleaner.

Fuzzy c-means clustering algorithm did not produce good performance at all the time due to dead centre and centre redundancy problems. As well as, they were also unable to avoid the centers from being trapped in the local minimum. The results obtained in Figures 1 (e) - 2 (e) for AML images and Figures 3 (e) - 4 (e) for ALL images show that the Moving k-means clustering algorithm produced better performance than the method based on Fuzzy c-means clustering algorithm. Furthermore, less of unwanted noise regions are left over using the combination of Moving k-means procedures as shown in Figure 1 (i) as compared to the combination of Fuzzy c-means procedures as shown in Figure 1 (h) and Figure 4 (h). The combination of Moving k-means procedures is more reliable with respect to noise. Hence, from Figure 2 (i) it shows that some of cytoplasm regions for blast in AML images can be detected as compared to Figure 2 (h). Generally, combination of Moving k-means procedures produced better performance as compared to Fuzzy c-means procedures, though some problems were still not totally avoided. For instance, Figure 1 (i) shows that unwanted noises can still be seen which is bigger in size (portions of red blood cells) and detected as part of the regions of interest.

Results of using pixel subtraction technique between manual segmented image in Figures 1 (c) - 4 (c) and resultant segmented image in Figures 1 (e) - 4 (e) are shown in Figures 1 (j) - 4 (j) and Figures 1 (k) - 4 (k) respectively. From the Figures 1 (j) - 4 (j) and Figures 1 (k) - 4 (k) are ghost images that represent the unsuccessfully segmented object appeared in original color.

For quantitative assessment, the resultant segmented image is compared to manual segmented image as reference. Manual segmented image were prepared by manually editing using Adobe Photoshop™. Figures 1 (c) – 4 (c) show example of manual segmented images. The segmentation performance of Moving k-means and Fuzzy c-means algorithm is evaluated by determine the percentage of accuracy, as in

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{9}
\]

where TP, TN, FP and FN are true positive, true negative, false positive and false negative.

The percentage of accuracy as in (9) is calculated based on comparison of the pixels that represent resultant segmented image and the pixels represent manual segmented image. If the test is correctly indicate the blasts (object interest), then the pixels that represent the blasts will be labeled as true positive (TP). If the test could correctly indicate the pixels that represent the background, then it will be labeled as true negative (TN). However, if the test could falsely indicate the blasts (object interest)
when truly is, then it labeled as false negative (FN). Finally, if the test could falsely indicate the pixels that represent the background, it will be labeled as false positive (FP). Figures 1 (h) – 4 (h) to Figures 1 (i) – 4 (i) show eight samples of segmented images used for testing. The number of data used for testing is 480,000 pixels that is equivalent to the size image which is 800 x 600. Table 1 shows the segmentation performance using Moving $k$-means and Fuzzy c-means algorithm. The results indicate that, segmentation using Moving $k$-means procedure is generally produced slightly better performance with the accuracy of 98.46% (AML1 type), 98.96% (AML2 type), 99.90% (ALL1 type) and 99.39 % (ALL2 type) respectively. As compared to Fuzzy c-means procedures the accuracy are 93.86% (AML1 type), 98.35% (AML2 type), 99.91% (ALL1 type) and 98.59 % (ALL2 type) respectively. Generally, this proposed method sustaining the size and shape of the blasts in the AML and ALL images. Additionally, the location of blast is successfully detected.

Table 1: Segmentation performance of the proposed technique

(a) AML type

<table>
<thead>
<tr>
<th>Type</th>
<th>Clustering algorithm</th>
<th>Accuracy (%)</th>
<th>TP (%)</th>
<th>TN (%)</th>
<th>FP (%)</th>
<th>FN (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML1</td>
<td>Fuzzy c-means</td>
<td>93.86</td>
<td>4.89</td>
<td>88.98</td>
<td>5.98</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Moving $k$-means</td>
<td><strong>98.46</strong></td>
<td>4.83</td>
<td>93.63</td>
<td>1.32</td>
<td>0.22</td>
</tr>
<tr>
<td>AML2</td>
<td>Fuzzy c-means</td>
<td>98.35</td>
<td>5.02</td>
<td>93.33</td>
<td>0.001</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>Moving $k$-means</td>
<td><strong>98.96</strong></td>
<td>5.68</td>
<td>93.28</td>
<td>0.05</td>
<td>0.99</td>
</tr>
</tbody>
</table>

(b) ALL type

<table>
<thead>
<tr>
<th>Type</th>
<th>Clustering algorithm</th>
<th>Accuracy (%)</th>
<th>TP (%)</th>
<th>TN (%)</th>
<th>FP (%)</th>
<th>FN (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL1</td>
<td>Fuzzy c-means</td>
<td><strong>99.91</strong></td>
<td>1.55</td>
<td>98.34</td>
<td>0.047</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>Moving $k$-means</td>
<td>99.90</td>
<td>1.52</td>
<td>98.38</td>
<td>0.027</td>
<td>0.077</td>
</tr>
<tr>
<td>ALL2</td>
<td>Fuzzy c-means</td>
<td>98.59</td>
<td>5.57</td>
<td>93.01</td>
<td>0.90</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Moving $k$-means</td>
<td><strong>99.39</strong></td>
<td>5.39</td>
<td>93.91</td>
<td>0.001</td>
<td>0.68</td>
</tr>
</tbody>
</table>

4. Conclusion

The current study proposed auto segmentation technique which consists of saturation component based on HSI color space, clustering algorithm, 7x7 pixels median filter and seed region growing algorithm. Two clustering algorithms namely Moving $k$-means and Fuzzy c-means are utilized. The performance of those clustering algorithms are compared and analyzed. The results show that method based on Moving $k$-means procedures yield better accuracy due to its capability in avoiding dead centre, pixels redundancy and getting trapped in local minima. The advantage of the proposed method is that the selection of the threshold for segmentation is done automatically. Furthermore, the blasts in acute leukemia blood samples are successfully segmented from its background and unwanted noises. The location of blast is successfully detected. Meanwhile, the shape for blast of acute leukemia has also been closely preserved. For the future work, the results from this paper can be used as the basis for extracting the other features from the acute leukemia blood samples. On the other hand, to
establish the capability and reliability of the proposed method more acute leukemia blood samples should be taken for further testing.

Acknowledgment

We would like to express our thanks to UniMAP and Malaysian Government for supporting this research in term of research grant. This research is funded under Fundamental Research Grant Scheme (Grant No. 9003 00129).

References

Fig. 1 Resulted images for AML 1, (a) Original image and resulted images after applying (b) saturation formula (c) manual segmentation (d) Fuzzy c-means algorithm (e) Moving k-means algorithm (f) and (g) 7x7 pixels median filter (h) and (i) seed region growing algorithm (j) and (k) Ghost Image.
Fig. 2 Resulted images for AML type (a) Original image and resulted images after applying (b) saturation formula (c) manual segmentation (d) Fuzzy $c$-means algorithm (e) Moving $k$-means algorithm (f) and (g) 7x7 pixels median filter (h) and (i) seed region growing algorithm (j) and (k) Ghost Image
Fig. 3 Resulted images for ALL 1 type (a) Original image and resulted images after applying (b) saturation formula (c) manual segmentation (d) Fuzzy c-means algorithm (e) Moving k-means algorithm (f) and (g) 7x7 pixels median filter (h) and (i) seed region growing algorithm (j) and (k) Ghost Image
Fig. 4 Resulted images for **ALL 2** type (a) Original image and resulted images after applying (b) saturation formula (c) manual segmentation (d) Fuzzy $c$-means algorithm (e) Moving $k$-means algorithm (f) and (g) 7x7 pixels median filter (h) and (i) seed region growing algorithm (j) and (k) Ghost Image