Computer aided system for lymphoblast classification to detect acute lymphoblastic leukemia

Syadia Nabilah Mohd Safuan, Mohd Razali Md Tomari, Wan Nurshazwani Wan Zakaria, Mohd Norzali Haji Mohd, Nor Surayahani Suriani
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:
Received Sep 26, 2018
Revised Nov 28, 2018
Accepted Jan 20, 2019

Keywords:
Acute lymphoblastic leukemia
Color space analysis
Support vector machine
White blood cell

ABSTRACT

Acute lymphoblastic leukemia (ALL) is a disease that is detected by the presence of lymphoblast cell. Basically, lymphoblast cell is the abnormal cell of lymphocyte which is one of the White Blood Cell (WBC) types. Early prevention is suggested as this disease can be fatal and caused death. Traditionally, ALL is detected by using manual analysis which is challenging and time consuming. It can also yield inaccurate result as it is highly dependent on the pathologist’s skills. Industry has come out with hematology counter which is fast, accurate and automated. However, these machines are costly and cannot be afforded by some countries. For that reason, Computer Aided System (CAS) will be a great help to the pathologist for assisting purposes and it also can act as second opinion for the pathologist. This system contains six main steps which are color space correction, WBC segmentation, post processing, clumped area extraction, feature extraction and lymphoblast classification. Firstly, color space correction is apply by using l*a*b* color space to standardize the image’s intensity. Next, WBC segmentation is made to prune out WBC region using color space analysis with Otsu thresholding. However, segmented image contains noises that need to be eliminated and it is accomplished by applying morphological filter with Connected Component Labelling (CCL). There is an overlapping WBC which need to be separated by using Watershed method to extract the individual cells. Next, feature extraction is made to collect the cell’s data to be fed into the classifier. Classifier used in this system to classify lymphoblast is Support Vector Machine (SVM) and this system is able to achieve 96.69% of accuracy.

Copyright © 2019 Institute of Advanced Engineering and Science. All rights reserved.

1. INTRODUCTION

Three main elements in a blood are Red Blood Cell (RBC), White Blood Cell (WBC) and platelet. We are exposed to many viruses and bacteria almost in every place. These situations can lead to sickness and death if human immune system is not strong to fight the viruses. In this case, WBC plays a very important role as its function is to fight bacteria and viruses in human body [1]. Having a good immune system is crucial to help fight sickness by monitoring the analysis of WBC in human body. Diseases such as HIV and Lymphoma can be diagnosed by low WBC count and diseases such as Leukemia and Anemia can be detected by a high WBC count. Leukemia is a very serious disease that can contribute to the percentage of death worldwide.
One of the leukemia types that is widely investigated is Acute Lymphoblast Leukemia (ALL). ALL is a very serious disease that is caused by the uncontrollable growth of abnormal lymphocyte called lymphoblast. Common symptoms of ALL is anemia, shortness of breath, fatigue, recurrent infection and unusual bleeding [2]. This disease also can lead to death if the patient left untreated which caused the lymphoblast to spread rapidly in the body [3]. ALL detection starts with the WBC identification and count in blood smear image. If the WBC count is low, bone marrow biopsy is performed to the patient [4]. Lymphoblast is differentiated by its shape and structure. There are shape irregularities in lymphoblast and it is also can be differentiated by the shape of its nucleus [5]. Analyzing WBC is a very crucial but tough task. Traditionally, WBC test is done manually by the pathologists that tend to yield inaccurate result. It is also highly dependent on the pathologist’s skill and might create confusion [2].

Computer Aided System (CAS) to detect ALL disease starts by identifying the presence of lymphoblast cell in blood smear image. One of the ways to detect and classify lymphoblast is by applying image processing technique in Computer Vision System. The system processes the image and detect WBC region and lymphoblast cell automatically. There are few main methods that are proposed to complete the process which are pre-processing, cell segmentation, post-processing, feature extraction and lymphoblast classification [6]. Purpose of post-processing step is to transform the original color of blood image to a more suitable level of intensity by using various types of method. There are many pre-processing techniques that is used in previous works such as gray-level conversion and contrast stretching [7], adaptive histogram equalization [8] and contrast adjustment [9]. Segmentation is reported to be one of the most crucial parts as the final result is highly dependent on this process [10]. This process can be divided into five main categories which are threshold-based, learning-bases, active-contour based, metaheuristic-based and saliency-based [11]. Threshold-based method is reported as the best method for the uniform image as blood cell image [12]. One of the works used combination of Otsu thresholding and Niblack binarization to segment the WBC [13]. Other than that, thresholding the saturation image also give a good result to eliminate variations in illumination [14]. Other than that, K-mean clustering is used to extract and prune out the area of WBC region color based segmentation [14, 15]. Next process is the post-processing step which is important to eliminate noises in the segmented image. After the noise has been distinguished, the image will go through feature extraction process that allows the system to take and calculated the cell nature.

There are many features that utilize in the previous works for the classification process. As presented in [16] and [17], features of fractal dimension, shape features which are contour signature and texture, and color features are extracted to classify lymphoblast in blood smear image and detect ALL. Other than that, shape features that contains elements such as area, perimeter, major axis, minor axis and solidity is used with addition of ratio between the area of cytoplasm and the nucleus and the number and structure of the core lobes [18]. Some of works used binary image to extract shape features, gray level image to extract mean gray level value and its standard deviation as explained in [3].

Lastly, the output of feature extraction will be the input for classification process. Classification is the process to categorize output based on its features and the ability of the system to classify the object accurately. One of the works used Support Vector Machine (SVM) to classify five types of WBC and it achieves 94.7% of lymphocytes classification by using spatial and spectral features [19]. Apart from that, SVM classifier is also used to classify lymphoblast cell to detect ALL in blood smear image. Accuracy of 93% and 93.57% has been achieved in [20] and [21] respectively. Other than SVM, Neural Network of Multi-Layer Perceptron – Back Propagation (MLP-BP) is also used for WBC classification and the work is able to achieve 96% of accuracy [22].

Combination of various features and segmentation method is used in this automated CAS for lymphoblast classification and ALL detection which is done by using thresholding with color space segmentation method, 16 features extraction and SVM classification.

2. RESEARCH METHOD

There are few main blocks in order to detect WBC region and classify lymphoblast cells. As depicted in Figure 1, it contains six main blocks which include color space correction, WBC segmentation, post processing, clumped cell area extraction, feature extraction and lastly, lymphoblast classification.
Firstly, color space correction is done to the original image which consist of RBC, WBC and platelet. Initially, there is an issue regarding the original image’s color intensity due to the different image acquisition condition. The idea of color space correction is to transform the original image’s color intensity to a standard color intensity. Next, WBC segmentation is done to prune out the area of WBC region only without any other substances in the image. However, the segmented image consists of noises and unwanted region that needed to be eliminated. It is solved by applying morphological filter to distinguish the noises. In the blood smear image itself, contains clumped and overlapping area of WBC region. Clumped area extraction step must be taken to individualize the WBC region. The individual or single WBC region image is obtained to extract the feature of the region of interest (ROI). Lastly, lymphoblast is classified from any other WBC types by using Support Vector Machine (SVM). The overall performance of the whole system is evaluated based on its ground truth data.

In this system, the original image or set of images is taken from IDB database [5]. It is the public blood smear image database that consists of three main elements which are RBC, WBC and platelet. In ALL-IDB(1) database, there are 108 images in total and the magnification factors are in the range of 300-500. These images are taken from the optical laboratory microscope with different lighting condition. As for this project, images with the same magnification factor, resolution and lighting condition is used. A constant multiplier needs to be tuned for other resolution and magnification factor images to achieve the right image ratio settings.

2.1. Color Space Correction

The color intensity of blood smear image that is taken from a microscope may varies due to the different acquisition condition such as setting and lighting of the microscope. In this paper, color space correction method is used as the basic idea is to standardize the color intensity of the image by transforming the current image’s intensity color to a targeted color representation.

Basically, this standardization process is done by matching RGB source and template images to \( I^*a^*b^* \) color representation. It starts by obtaining the values of mean and standard deviation of \( I^*a^*b^* \) of both template images and source image. Next, the values of each \( I^*a^*b^* \) of mean and standard deviation is substituted in the Equation (1) below. \( Im \) represents the single color representation of \( I \), \( a \) and \( b \) which will finally leads to corrected value of \( I \) a and b. These values are then applied to the original image and change its color intensity thus standardize the final image result.

\[
Im_{corrected} = \frac{Im_{source} - \mu_{source}}{\sigma_{source}} + \mu_{template}
\]

(1)

However, in our implementation, instead of using specific image itself as the target and match with the source image, four template images are used as the target image as it qualitatively produces good color representation as shown in Figure 2 below. These images were taken from the device in our laboratory and were chosen as a template image as it highlights the WBC region better. The images taken from other devices may varies in terms of the color distribution and have the tendency to decrease the system performance. For that reason, it is important to standardize the color distribution before proceeding with segmentation process. For each template image, the value of mean and standard deviation is taken and average value of mean and standard deviation for four template images are calculated.
The obtained average mean values of l, a and b are 81.9863, 4.5579 and 1.2176 respectively while average standard deviation values are 4.7984, 4.7248 and 3.0947. The color intensity of source image or original image is matched to these values and the comparison between original color and color corrected image is as depicted in Figure 3. It can clearly be seen that the original image has been transformed to a color that are close to the template image. Furthermore, the image characteristic and structure will not be affected by this process.

2.2. WBC Identification And Segmentation

This section explains about the most important part of the project which is the WBC segmentation and extraction. Basically, segmentation is a detection process where WBC region is prune out from the blood smear image. Presence of other particles such as RBC, platelet and background need to be minimized to achieve high WBC detection accuracy. However, this is a challenging task as WBC itself contains two elements which are cytoplasm and nucleus as described in Figure 4. Nucleus is the inner part while cytoplasm is the outer part and both parts can be differentiated by their color.

As for this process, Otsu thresholding is applied on color space analysis of RGB, HSV and CMYK. Figure 5 depicts the result of segmentation for some color space and it is analyzed to determine the best color representation that is capable to prune out WBC nucleus and cytoplasm. These results are compared to the ground truth image as shown in Figure 6 to determine the closest color analysis result that can be used for...
nucleus and cytoplasm detection. It can be seen that the best three for nucleus segmentation with a total elimination of background and RBC can be taken from the single color band analysis of G of RGB, S of HSV and C of CMYK.

![Color channel segmentation](image1.png)

*Figure 5. Result of color channel segmentation of hue (H), saturation (S), value (V) and yellow (Y)*

![Ground truth data](image2.png)

*Figure 6. Sample of ground truth data of nucleus and cytoplasm*

However, there are no single color space analysis that can match the cytoplasm detection ground truth. In this case, other approach of detecting WBC cytoplasm is taken which is the combination of two single color bands. H of HSV and Y of CMYK show the largest area of WBC detection which include most of the cytoplasm area as well. Next, this color band analysis is combined by subtracting Y from H. The result of H-Y subtraction is depicted in Figure 7 and it can be seen that the WBC cytoplasm has been fully detected which include the nucleus area. The RBC and background is also eliminated from the image, left only cytoplasm area.

![H-Y subtraction](image3.png)

*Figure 7. H-Y subtracted image*
2.3. Post-Processing

Result of segmented image for both nucleus and cytoplasm contains noises and unwanted region that need to be distinguished. It is very crucial to get rid of the unwanted region as it will definitely lessen the accuracy of whole system performance. In this paper, post-processing step is done by applying morphological filter which includes erosion and dilation. Erosion is used to shrunken and erode the binary object in the image while dilation used to thicken and grow the region. Both method is applied once for each segmented image. The result of morphological filter application is as shown in Figure 8. However, getting rid of the unwanted region is not an easy task. There are still some noises left and for this case, Connected Component Labelling (CCL) of 150 is applied. Basic idea of CCL is to remove the certain value of pixel in the image. In our implementation, any object or region that is below than 150 pixels will be removed from the image. This process will not affect any other region that is more than 150. As depicted in Figure 8, the structure, shape and size of that region will remain unchanged.

![Figure 8. (a) Image before post processing (b) Result of post processing](image)

2.4. Clumped Area Extraction

In ALL-IDB (1) database, there are some WBC region that is clumped and overlapping with each other. Individual cell of WBC need to be extracted before feature extraction can be taken for lymphoblast classification. Each individual lymphoblast and non-lymphoblast cell have to be separated from one and another.

In order to separate these clumped cells, firstly, clumped region area is defined by calculating the average area of each individual cell. CCL method is applied once again to prune out only the clumped cell area. This time, CCL value used is 750 which means that any object that is above 750 is considered as an overlapping cell region. As a result, only clumped cell area is remained in the image and the individual cells are removed. Next, separation process to extract the individual cell is done by using Watershed segmentation to the distance transform. Watershed’s basic idea is to detect the boundary line of the connected region and split them into individual region and it is based on the distance between the pixel’s boundaries.

2.5. Feature Extraction

Feature extraction is a technique of redefining a large set of redundant data into a set of features of reduced dimension. As lymphoblast is differentiated by its shape and structure, important features was extracted based on geometric features, texture features and color information. For the first section which is the shape feature, it contains of four elements. Elements that have been taken is area, compactness, convex area, and solidity. These features are obtained by the binary sub-image as shown in Figure 9.

![Figure 9. Binary nucleus sub-image](image)
Next, second section is texture feature. Texture feature measurement is applied on the gray-level sub-image as shown in Figure 10. First part is based on Gray Level Co-occurrence Matrices (GLCM) which consists of four elements which are homogeneity, energy, correlation and contrast. After that, the second part contains another eight elements which are mean, standard deviation, entropy, RMS, variance, smoothness, kurtosis, skewness and Inverse Different Moment (IDM). There are 16 features in total to be extracted in each individual cell for classification process later on.

![Gray-level sub-image](image)

**Figure 10. Gray-level sub-image**

### 2.6. Lymphoblast Classification

Classification is a process to assign the unknown data to the one of the labelled of known class. In this paper, classification is focused on two classes which are lymphoblast and non-lymphoblast class. In order to classify the lymphoblast and non-lymphoblast, Support Vector Machine (SVM) is employed to the system. SVM is reported to have a stable performance with less fluctuation compared to one of the popular classifiers, Neural Network [23]. SVM works by creating a hyperplane in the F space (input space) that has maximum margin separation. In this work, there are two class classifiers that optimize the margin [16].

### 2.7. System Performance Assessment

After all the method has been applied in the system for each blood smear image, the system performance is evaluated based on quantitative analysis. It is divided into two main parts which are the performance of WBC counting and the performance of WBC identification.

#### 2.7.1 WBC Identification

Performance of WBC identification is crucial to make sure that the WBC region is correctly identified. It is conducted based on the comparison between automated segmentation result image and the ground truth image. Four important parameters are taken from the difference between segmented image and the ground truth data. The parameters are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Basically, TP defines WBC region that is correctly identified as WBC while TN explains non-WBC region that is detected as non-WBC region. Other than that, FP states the region of non-WBC that is detected as WBC while FN tells the WBC region that is detected as non-WBC. After all these values are obtained, the accuracy, specificity and sensitivity is calculated by substituting these four parameters in the Equation (3), (4) and (5).

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3) \\
\text{Specificity} = \frac{TN}{TN+FP} \quad (4) \\
\text{Sensitivity} = \frac{TP}{TP+FN} \quad (5)
\]

#### 2.7.2 WBC Counting

The second part of system performance assessment is the WBC counting accuracy. This assessment needs to be evaluated to practically know the best segmentation method and the best color space analysis that can detect WBC correctly. WBC counting process is quite a challenging task as the irregular shape of WBC...
WBC counting performance is evaluated by applying Circle Hough Transform (CHT) method. It works by calculating the range of radius that has been set in the algorithm. Firstly, it detects the circle region and starts predicting the radius of the circle. It can be defined by the Equation (6) below where a and b is the coordinate of the circle and r is the predicted radius. It detects the edge of the circle by drawing and circling the region.

\[(x-a)^2 + (y-b)^2 = r^2\]  

(6)

### 3. RESULTS AND ANALYSIS

This system is developed using Matlab software by applying the image processing toolbox. There are 108 images in total in the ALL-IDB (1) database with different magnification factor and resolution. However, in our implementation, images with same magnification factor and same resolution which is 1712x1368 is only considered in order to maintain the accuracy and achieve consistent result. While in ALL-IDB (2) database, there are 260 individual cells which consist of lymphoblast and non-lymphoblast cell.

#### 3.1. WBC Counting

Firstly, WBC segmentation and identification accuracy is evaluated by counting the WBC region in binary blood smear image. It has been mentioned before, qualitatively, the best three for nucleus segmentation with a total elimination of background and RBC can be taken from the single color band analysis of G of RGB, S of HSV and C of CMYK. While for the cytoplasm detection, H-Y subtraction method is used. All these segmentation results of G, S, C and H-Y is taken and the WBC counting for each method is evaluated. Next, the average accuracy for 30 blood smear images is calculated and shown in Table 1. The counting accuracy is for each method is compared and it can be seen that segmentation using S of HSV produces the highest result of WBC counting. Nucleus segmentation S of HSV achieves 96.92% while cytoplasm segmentation H-Y subtract gives accuracy of 40.72%.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>ACCURACY (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>96.92</td>
</tr>
<tr>
<td>C</td>
<td>95.13</td>
</tr>
<tr>
<td>G</td>
<td>59.39</td>
</tr>
<tr>
<td>H-Y</td>
<td>40.72</td>
</tr>
</tbody>
</table>

Table 1. WBC Counting of S, C, G and H-Y

There is a huge difference between the best nucleus segmentation and cytoplasm segmentation counting accuracy result. This happens because of the irregular shape of cytoplasm area. As a result, segmented area of single cytoplasm is miscalculated as two or three WBC cells because of its large area compared to nucleus.

In the proposed framework section, the first main block is color space correction. The motivation to include this step in the system is explained previously. However, to make the justification stronger, effect of color space correction to the WBC counting is studied. Basically, the result of WBC counting performance with and without color space correction is compared. As for this purpose, the comparison is made using the highest accuracy result obtained from Table 1 for nucleus segmentation which is S of HSV and cytoplasm counting accuracy. The comparison is depicted in Table 2 which clearly shows that counting accuracy without the presence of color space correction process is deteriorate for both nucleus and cytoplasm segmentation. It is clearly can be seen that color space correction process is highly needed for the system in order to achieve high WBC counting accuracy result.

<table>
<thead>
<tr>
<th>WBC Segmentation</th>
<th>With Color Space Correction</th>
<th>Without Color Space Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus (S)</td>
<td>96.92 %</td>
<td>93.55 %</td>
</tr>
<tr>
<td>Cytoplasm (H-Y)</td>
<td>40.72 %</td>
<td>17.55 %</td>
</tr>
</tbody>
</table>

Table 2. Color Space Correction Comparison
3.2. WBC Identification

In this section, the performance of WBC identification is evaluated based on its ability to detect WBC region accurately. Based on the WBC counting accuracy, S of HSV color analysis is selected as it provides the highest counting accuracy and H-Y subtract from the cytoplasm based segmentation. Average percentage of accuracy, specificity and sensitivity for 30 images is calculated for both nucleus and cytoplasm area detection as shown in Table 3. Nucleus area detection provides higher accuracy and specificity result. Overall, nucleus based result is more satisfying for both counting accuracy and detection accuracy.

<table>
<thead>
<tr>
<th>WBC Identification</th>
<th>Accuracy (%)</th>
<th>Specificity (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus (S)</td>
<td>98.87</td>
<td>96.87</td>
<td>99.10</td>
</tr>
<tr>
<td>Cytoplasm (H-Y)</td>
<td>74.12</td>
<td>65.32</td>
<td>99.87</td>
</tr>
</tbody>
</table>

3.3. Lymphoblast Classification

From the sub-images that contain the individual WBC region, the nature of each cells are taken along with their labels. In this paper, there are only two classes which are lymphoblast and non-lymphoblast. Total individual WBC sub-image is 242 images. Basically, features matric with a size of 16 x 242 and classification vector of a size 1 x 242 is created.

Classifier that is used in this system is Support Vector Machine (SVM) as it can perform a reliable result with small amount of training data. SVM creates a hyperplane in the input space to classify two categories. In order to calculate the accuracy, 500 iterations are used and the data is trained by applying linear kernel function. The accuracy obtained for lymphoblast classification is 96.69%.

4. CONCLUSION

In this work, we have proposed an automated system to identify and classify lymphoblast cell in blood smear image for Acute Lymphoblastic Leukemia (ALL) detection by developing computer vision system using image processing purposes. There are many methods that is proposed by other works to classify lymphoblast and almost all the works used the same main block of steps which are pre-processing, ROI segmentation, post-processing, feature extraction and classification. The result of color space correction shows the RBC region slightly faded. Next, WBC segmentation is done by using color space analysis along with Otsu thresholding. Noise elimination by applying morphological filter is used with addition of CCL method. After the unwanted region is completely distinguished, individual WBC region is extracted by using watershed segmentation. As a result, individual WBC region sub-image is created to obtain the nature features of the cell. Selected features are chosen and obtained for each sub-image. Lastly, this data of features is used as input for classification process using SVM classifier. Result shows that this system can classify lymphoblast correctly and accurately.

Future development than can be expanded in this work is to vary the dimension of input image as this system only applicable for image with same resolution and magnification factor. Next, it can also be developed on mobile device to make it more portable and a compact machine.

ACKNOWLEDGEMENTS

The authors would like to thank to Ministry of Education (MOE) and Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research under Fundamental Research Grant Scheme (Vot. no. 1582).

REFERENCES


BIOGRAPHIES OF AUTHORS

Syadia Nabilah Mohd Safuan received B.Eng (2016) in Electronics Engineering from Universiti Tun Hussein Onn Malaysia and is currently doing master in Electrical Engineering at Universiti Tun Hussein Onn Malaysia. Her current research interest includes image processing and computer vision system. She is author and co-author of several journal papers and conference proceedings.
Computer aided system for lymphoblast classification to detect acute… (Syadia Nabilah Mohd Safuan)