Article

Geometric Feature Extraction for Identification and Classification of Overlapping Cells for Leukaemia

Siew Ming Kiu and Yin Chai Wang *

Faculty Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan 94300, Sarawak, Malaysia; eve_kiu@hotmail.com
* Correspondence: ycwang@unimas.my

Abstract: This paper describes the study of overlapping leukaemia cells based on geometric features for identification and classification. Geometric features of blood cells are proposed to identify and classify overlapping cells into groups based on different overlapping degrees and the number of overlapped cells. In the proposed method, the percentage of average accuracy for identifying overlapping cells reached 98 percent. The proposed method can segment white blood cells from the overlapping cells with an overlapping degree of 70 percent. Improved Watershed Algorithm successfully increased 36.89 percent of accuracy in WBC segmentation. It reduced 46.12 percent of the over-segmentation problem. Tests of cell counting are conducted in the two stages, which are before and after the process of identification and classification of overlapping cells. The average percentage of total cell count is 83.31 percent, the average percentage of WBC counting is 84.8 percent, and the average percentage of RBC counting is 60.55 percent. The proposed method is efficient in identifying and classifying overlapping cells for increasing the accuracy of cell counting.

Keywords: image processing approach; overlapping cells; WBC; RBC; geometric feature extraction; identification; classification; leukaemia; overlapping degree; Improved Watershed Algorithm

1. Introduction

Image processing is methods or algorithms that perform some operations on an image to obtain useful information. It is now routinely used by individuals who have access to computers and digital cameras. In a clinical laboratory, complete blood cell count is used to diagnose diseases by counting a number of cells and identification of anomalies. For instance, leukaemia can be diagnosed by counting the abnormal number of WBC in the patient’s blood. Blood cell count is executed by the image segmentation technique in image processing. The characteristics of blood cells are used for cell identification, and blood cells are counted after the cells are segmented by the segmentation method. Accurate blood cell count is significant for the clinical diagnosis and quantitative analysis. Previously developed automated cell counting by others, which focus more on the classification of cell type, have lower accuracy when there are highly overlapping cells and WBC with irregular shapes and sizes during the counting process. For example, Hiremath et al., (2010) [1] classified WBCs into three types of WBCs, which are lymphocyte, monocyte, and neutrophil. In the case study of Mojtaba Taherisadr et al., (2013) [2], red blood cells were classified into twelve categories, such as normal cell, elongated cell, target cell, helmet cell, microcyte, spherocyte, and sickle cell. Besides, overlapping cells is another issue faced by previous researchers that affect the accuracy of CBC. Many researchers also mentioned this problem in their studies, such as Mao et al., (2003) [3], Nguyen et al., (2010) [4], Khan and Maruf (2013) [5], Fan et al., (2013) [6], Lu et al., (2015) [7], and Phoulady et al., (2016) [8]. Hence, the overlapping case becomes the main problem that needs to be solved in this study. A geometric feature extraction method is proposed to identify and classify the overlapping cells to improve the precision of blood cell count. In the process of splitting overlapping...
cells, the most commonly used method is Watershed Transform Algorithm. The Watershed Transform Algorithm has the problem of under-segmentation and over-segmentation when there is the presence of cells with a high overlapping degree. This phenomenon was mentioned in the studies of Lim et al., (2015) [9], Liu et al., (2015) [10], Khan et al., (2013) [5], Sharif et al., (2012) [11], and Li et al., (2010) [12]. Furthermore, over-segmentation decreases the accuracy of blood cell count, which was mentioned by Roerdink and Meijster (2000) [13], Bieniecki (2004) [14], Bala (2012) [15], Amoda and Kulkarni (2013) [16], and Arslan et al., (2014) [17]. The Improved Watershed Algorithm is recommended to replace the Traditional Watershed Transform Algorithm, which was widely chosen by previous researchers to carry out the work to reduce the possibilities of over-segmentation and accelerate the accuracy of blood counting in the post-processing. This is especially more obvious when there are more highly overlapped cells and WBC with irregular shapes.

2. Materials and Methods

2.1. Database Benchmark

One hundred and three microscopic images used in this study were collected from online image databases, namely ASH Image Bank and Medical Stock Image. Another 164 real microscopic images were edited by using Adobe Photoshop CS6 to form different magnifications ranging from 100× to 500× and saved in JPEG format. Additionally, there are 100 artificial overlapping cells images that were formed by cropping off the single WBC from real microscopic images to form additional overlapping cases with a different number of cells. All artificial JPEG images were created in 100× microscope magnifications with a dimension of 450 × 450 pixels and processed using MATLAB R2014a.

2.2. Geometric Feature Extraction

Geometric feature extraction, which is proposed as the solution to identify and classify the overlapping cell into a single cell for cell counting purposes, is discussed. The general method for object counting contains four steps, namely, image acquisition, image enhancement, image segmentation, and object counting (Figure 1).

![Figure 1. Framework of object counting using image processing.](image)

An additional step in between image segmentation and object counting, which is feature extraction, was added to the framework to increase the accuracy of object counting. The complete proposed method is shown in Figure 2. The additional step is implemented with the writer’s source code.

This method starts with image acquisition. Required images are acquired by capturing images through downloading images from an online public image library. Collected images are processed for further enhancement.

Image enhancement is a pre-processing step for post-processing, such as image segmentation, feature extraction, cell segmentation, and counting. The image enhancement techniques of grey level transformation are used to eliminate the hue and saturation information while retaining luminance. Furthermore, noise removal is applied for the improvement of the acquired image.
Image segmentation is the process of dividing an image into multiple parts, such as foreground and background. A Triangle method proposed by Zack et al., (1977) [18] was used in the proposed method as an Automatic Threshold Method. This technique is applied to the green component of the image, as the image has a value representing the greatest differences between minimum and maximum intensity levels. In Figure 3, which is an example of the triangle method, a line is constructed between the highest value, $b_{\text{max}}$, and lowest value, $b_{\text{min}}$, where $p = 0$, indicates the values of the grey levels where the histogram $h[b]$ reaches its maximum and minimum, respectively. Distance, $d$, in between the marked line and histogram, $h[b]$, which is in between $b_{\text{min}}$ and $b_{\text{max}}$, is then calculated. The intensity value, where the distance $d$ reaches its maximum, defines the threshold value (Threshold = $b_0$). This algorithm was proposed by Putzu and Ruberto (2013) [19], and it is a more effective way to threshold the objects from the background in a short period of time. It is not similar to the global thresholding method, which spends time selecting suitable threshold values for each image, and it wastes time if there is a large number of images that need to be processed.

Feature extraction was carried out to identify overlapping cells. Before feature extraction, each blob of a cell was detected and labeled with a unique ID by using the bwlabel function. The function of regionprops was used to measure each blob’s properties. The geometric feature is the most simple and fastest way to identify overlapping cells, and it
is preferred and suggested by many researchers in their research, such as Fatichah et al., (2014) [20], and Reta et al., (2010) [21]. Geometric features of the area and roundness were determined for the identification of a single cell from the overlapping cell. The rules used to detect overlapping cells are single RBC that contain a smaller region pixel than WBC; WBC containing a roundness value greater than 0.8 and larger than RBC; and overlapping cells containing roundness smaller than 0.8 and a size larger than RBC or WBC. After the overlapping cell is identified, overlapping cells pass through the process of classification. The classification steps are shown in Figure 4.

1. Read the binary image of overlapping cell
2. Compute the Euclidean distance transform of a binary image
3. Compute the Watershed Transform of D
4. Create the watershed ridge lines in white colour, corresponding to $Ld == 0$. Hence, use it to segment the binary image by changing the corresponding pixels in the background
5. Call imextendedmin to filter out the local minima (small spot)
6. Call imimposemin to modify the distance transform in order to make sure no minima occur at the filtered-out location
7. Repeat the watershed steps from step 8 to step 10
8. Label the binary image of overlapping cell after splitting the overlapping cell.
9. Measure the region properties of each single cell
10. Use bwconncomp to find the connected components in bw3 of binary image
11. Use the number returned by bwconncomp function to know the number of cells overlaps in each set of overlapping cell
12. Start an array to loop each set of overlapping cells for identifying the number of cells overlap
13. If the number returned by step 11 is less than 3, it defines the set is a double cell overlapping
14. Else it defines the set is multiple cell overlapping
15. End array
16. End

Figure 4. The step of classifying overlapping cells into the number of cells overlapped.

The overlapping degree of overlapping cells is calculated by the addition rule in the process of classification. They are classified into four groups based on overlapping degrees (the percentage of overlapped) such as touching (intersection of two cells is 0), partial overlapping (intersection of two cells is more than 0 percent but less than 35 percent), medium overlapping (intersection of two cells is more than 35 percent but less than 60 percent), and complex overlapping (intersection of two cells is more than 60 percent but less than 80 percent). The number of cells overlapped is calculated by applying the Improved Watershed Algorithm to split the overlapping cell. After the number of cells overlapped is obtained, they are classified into different groups based on the number of cells overlapped. The groups are non-overlapping ($n = 1$), double overlapping ($n = 2$), and multiple overlapping ($n > 2$).

Cell segmentation is processed by the Improved Watershed Algorithm after cell classification. Improved watershed Algorithm removed the presence of “noise spots” in the region of connected components, which is local minima and surrounded by a watershed
region in the image. The watershed ridge line is formed, and the connected cell is split according to the watershed region after implementing the Watershed Transform Algorithm in the binary image. Each local minima becomes a catchment basin and causes the over-segmentation problem in the image. Therefore, the removal of noise spots is a necessary task to reduce over-segmentation by implementing the Improved Watershed Algorithm. After the overlapping cell is split into a single cell, cell counting is carried out on the segmented cell to calculate the number of RBC and WBC. The step of counting the cell quantity is shown in Figure 5.

1. Start
2. Set an array to calculate area pixel of each cell to group the WBC and RBC
3. It defines as WBC if area pixel is more than 3000 pixels, otherwise RBC
4. Count the number of each single cell before and after overlapping cell is splitting
5. End

Figure 5. The step of counting the number of every single cell.

3. Results and Discussion

The experimental results of the identification and classification of an overlapping cell are discussed. A comparison between the average accuracy in the manual method and the proposed method was made, and the performance of the proposed method for overlapping cells was evaluated and examined. An experiment was carried out to assess the ability of the proposed method to perform segmentation and counting with higher accuracy.

3.1. Experiment I: Overlapping Cells Identification

A total of 103 samples of the peripheral blood smear image of the leukaemia blood sample were in this experiment to evaluate the performance of the proposed method for identifying the overlapping cells. The results of the proposed method were compared with manual counting to evaluate the accuracy of identifying the number of cells in overlapping cells. In this experiment, accuracy, precision, false positives rate, and true positives rate were used as the metrics to evaluate the overlapped cell identification.

The accuracy of overlapping cell identification was carried out by comparing the segmentation results of auto count by the proposed method and manual count by a human expert. This metric measurement was by a few researchers in their studies, such as Sadeghian et al., (2009) [22], Putzu and Di Ruberto (2013) [19], and Fatichah et al., (2014) [20]. All of them used this metric to obtain the average accuracy for their performance of the proposed method. The equation of accuracy is shown as follow:

$$\text{Accuracy} = \left( \frac{\text{correctly predicted class}}{\text{total testing class}} \right) \times 100 \tag{1}$$

Figure 6 shows the sample image of overlapping cells identification, which is segmented manually by the proposed method. Manual segmentation marks a red-coloured box around overlapping cells manually, as shown in Figure 6b. The result in Figure 6c shows the successful identification of the same number of overlapping cells, which is carried out by the proposed method and compared with the manual methods. The feature of morphological was applied in the proposed method to identify overlapping cells. The result of manual segmented was used as a true value for calculation the accuracy of overlapping cells identification, which was performed by the proposed method.
The average percentage precision of overlapping cells identification was evaluated. The precision was calculated as the ratio of the number of true positives to the sum of true positives and false positives. True positives were calculated as the number of pixels (overlapping cells) that were correctly identified as overlapping cells. False positives were calculated as the number of pixels (single cell) that were incorrectly identified as overlapping cells. The equation is shown as follows:

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

(2)

The average accuracy and precision reached 98 percent and 92.26 percent, respectively. Thus, the measured value of the proposed method in WBC identification is 92.26 percent, close to the other measured value. Therefore, the proposed method for overlapped cell identification is generally accurate and precise.

3.2. Experiment II: Overlapping Cells Classification by Overlapping Degree

A total of 164 artificial JPEG images of overlapping cells with dimensions 450 × 450 pixels and with a different overlapping degree were created in this experiment. The overlapping degree of overlapping cells was calculated by the addition rule and compared with manual calculation. The relationship of the manual method and proposed method for accurate calculation of the overlapping degree of overlapping cells is shown in Figure 7.

![Figure 6](image-url)

**Figure 6.** Sample of overlapping cells identification. (a) Original image (John, 2011). (b) Overlapping cells manually segmented. (c) Overlapping cells segmented by the proposed method.

Based on the results, our proposed method successfully identified overlapping cells with overlapping degrees by more than 60 percent. The figure shows both results of overlapping degree contain slight difference within the range of 2 percent when two of the cells are overlapping below 60 percent. Moreover, the difference between manual and automatic calculation is within the range of 4 percent to 6 percent difference if there are cells overlapping more than 60 percent. The results are considered generally accurate and validated. It can be proven that the output of the automatic calculation is generally close to the output of manual calculation by using Person Correction Coefficient \((r)\). The equation of Person Correction Coefficient \((r)\) is shown below, where \(n\) is the number of images, \(\sum xy\) is the sum of the products of paired scores, \(\sum x\) is the sum of scores by \(x\) (manual method), and \(\sum y\) is the sum of scores by \(y\) (proposed method).

\[ r = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} \]

(3)
\[ \sum y \] is the sum of scores by \( y \) (proposed method), \( \sum x^2 \) is the sum of squared \( x \) scores, and \( \sum y^2 \) is the sum of squared \( y \) scores.

\[
    r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}
\] (3)

Table 1 shows the data for two variables and the answer validation of the manual and the proposed method. By using Equation (3), all the correct input values were plugged into the equation to obtain the result of the coefficient value, which is 0.98769. Based on the result, it shows that the variables have a strong positive correlation. Therefore, it can be said that the output between the manual and the proposed method is highly similar.

<table>
<thead>
<tr>
<th>Image</th>
<th>X (Manual)</th>
<th>Y (Auto)</th>
<th>X</th>
<th>X^2</th>
<th>Y^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.3</td>
<td>7.5</td>
<td>69.75</td>
<td>86.49</td>
<td>56.25</td>
</tr>
<tr>
<td>2</td>
<td>25.2</td>
<td>23.5</td>
<td>592.2</td>
<td>635.04</td>
<td>552.25</td>
</tr>
<tr>
<td>3</td>
<td>32.5</td>
<td>31</td>
<td>1007.5</td>
<td>1056.25</td>
<td>961</td>
</tr>
<tr>
<td>4</td>
<td>42.4</td>
<td>41.5</td>
<td>1799.6</td>
<td>1797.76</td>
<td>1722.25</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>37.5</td>
<td>3507.5</td>
<td>3721</td>
<td>3306.25</td>
</tr>
<tr>
<td>6</td>
<td>10.8</td>
<td>6.3</td>
<td>68.04</td>
<td>116.64</td>
<td>39.69</td>
</tr>
<tr>
<td>7</td>
<td>21.6</td>
<td>17.2</td>
<td>371.52</td>
<td>466.56</td>
<td>295.84</td>
</tr>
<tr>
<td>8</td>
<td>35.6</td>
<td>35</td>
<td>1246</td>
<td>1267.36</td>
<td>1225</td>
</tr>
<tr>
<td>9</td>
<td>52.1</td>
<td>42.2</td>
<td>2198.62</td>
<td>2714.41</td>
<td>1780.84</td>
</tr>
<tr>
<td>10</td>
<td>53.6</td>
<td>47.7</td>
<td>2556.72</td>
<td>2872.96</td>
<td>2275.29</td>
</tr>
<tr>
<td>Total</td>
<td>344.1</td>
<td>309.4</td>
<td>13,377.45</td>
<td>14,734.47</td>
<td>12,214.66</td>
</tr>
</tbody>
</table>

3.3. Experiment III: WBC Segmentation by Improved Watershed Algorithm

A total of 51 numbers of the peripheral blood smear image of the leukaemia blood samples were selected based on the requirement of high overlapping for this experiment to evaluate the performance of the Improved Watershed Algorithm to split the overlapping cells into the correct number of cells. The results of the Improved Watershed Algorithm were compared with Watershed Transform Algorithm to evaluate the accuracy. The results of splitting overlapping cells by Watershed Transform Algorithm and Improved Watershed Algorithm are shown in Figure 8.

Figure 8. The graph evaluates the performance of different splitting algorithms for WBC Segmentation.

The number of WBC segmentation of Watershed Transform Algorithm managed to obtain 51 cells out of the total number of 206 during the performed experiment. Improved Watershed Algorithm, vice-versa, obtained 127 out of the total number of 206 cells. In summary, the percentage of successful WBC segmentation by Watershed Transform Algorithm...
is 24.76 percent, whereas Improved Watershed Algorithm is 61.65 percent. The percentage of WBC segmentation increased by 36.89 percent when comparing the performance of the Improved Watershed Algorithm with the Traditional Watershed Transform Algorithm. Improved Watershed Algorithm reduced over-segmentation cases by 46.12 percent when compared to Traditional Watershed Transform Algorithm. However, the under-segmentation problem increased by 9.23 percent. All the results had shown in Table 2.

<table>
<thead>
<tr>
<th>WBC Segmentation</th>
<th>Total Number of Cells</th>
<th>Success Segmented Number</th>
<th>Success Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watershed Transform Method</td>
<td>206</td>
<td>51</td>
<td>24.76</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>206</td>
<td>127</td>
<td>61.65</td>
</tr>
</tbody>
</table>

3.4. Experiment IV: Cell Counting by Proposed Method

A total of 40 images with serious overlapping and WBC with irregular shapes were selected from 103 actual peripheral blood smear images as sample images in this experiment. The algorithm was run to carry out cell counting works on these 40 images. The experiment was carried out to determine the accuracy of the proposed method in cell counting by comparing the accuracy between the stage before and after the process of identification and classification for overlapping cells.

In Figure 9, it is clearly observed that the result of cell counting after applying the proposed method to identify and classify the overlapping cells is higher than the result of cell counting without the identification and classification of the overlapping cells. The result of the experiment shows that the average percentage of accuracy of the total cell count is 83.31 percent after the application of the proposed method. It increased 44.67 percent from 38.64 percent. For the WBC count, the result shows that the average percentage of accuracy is 84.8 percent, which is an increment of 23.74 percent from 61.06 percent. For the RBC count, the result shows that the average percentage is 60.55 percent, which increased to 32.55 percent from 28 percent. Therefore, this experiment achieved the objective of identifying the overlapping cells to improve the overall accuracy of cell counting.

![Figure 9](image-url)  
Figure 9. Comparing the result of cell counting between the stage before and after the process of identification and classification of overlapping cells.

4. Conclusions and Future Work

In conclusion, an algorithm was proposed to improve the accuracy of cell counting. It was achieved by the application of procedures, such as geometric feature extraction for the identification and classification of overlapping cells. The significant contribution of this
study is the successful improvement of the accuracy of cell counting with the proposed method, which can detect and perform for high overlapping cells with overlapping degrees higher than 60 percent. Improved Watershed Algorithm can reduce the over-segmentation problem by 46.12 percent. The success rate of the Improved Watershed Algorithm for WBC segmentation increased by 36.89 percent compared to the Traditional Watershed Transform Algorithm.

However, there are some limitations to be improved in the future. The accuracy of detection and calculation of overlapping cells with overlapping degrees higher than 80 percent and overlapping cells with more than six cells involved should be further improved. Additionally, post-processing should be included in the Improved Watershed Transform Algorithm to prevent the merging of some relevant watershed lines and prevent the occurrence under segmentation.

Author Contributions: Conceptualisation, S.M.K. and Y.C.W.; methodology, S.M.K. and Y.C.W.; software, S.M.K.; validation, Y.C.W.; formal analysis, S.M.K.; investigation, S.M.K.; resources, Y.C.W.; writing—original draft preparation, S.M.K. and Y.C.W.; writing—review and editing, Y.C.W.; supervision, Y.C.W.; project administration, S.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References


