Multiple Cells Extraction Algorithm on Peripheral Blood Smear Images Using Modified Significance Linked Connected Component Analysis

Chen-Chung Liu
National Chin-Yi University of Technology, Department of Electronic Engineering
ccl@ncut.edu.tw

Chun-Hao Fang
National Chin-Yi University of Technology, Department of Electronic Engineering
qqcandy111@gmail.com

Abstract

Many diseases such as Cancer, Huntington’s disease, Malaria, Multiple Sclerosis, and Myalgic Encephalomyelitis are mainly diagnosed by the appearances include size, shape, color and the double-concave dish size of red blood cells in the blood smear. The automatically evaluating process in the appearances from the blood smear images is high importance for rapid and accurate diseases diagnosis. The proposed algorithm combines several schemes to segment red blood cells in Giemsa stained blood slides and count the appearances of red blood cells in the blood smear image for advance applications. Experiment results show that the proposed algorithm not only to provide the above appearances of red blood cells quickly and accurately, but also provides the red blood cells and double-concave dishes’ median, average, standard deviation for physicians in different areas to diagnose different diseases.

Keywords: blood cells, blood slides, segment.

1. Introduction

Recently, more and more doctors and pathologists adopt digital medical image processing techniques and systems to help themselves to efficiently and accurately diagnose diseases. One of the most popularly used technique for diagnosing many diseases is visual microscopical evaluation of Giemsa stained blood smears. This process consists of manually counting the blood cells in a slide. The manual analysis of slides is laborious, time consuming, and requires experts [1]. Moreover, the accuracy of the final diagnosis is almost dependent on the operators’ skill, experience, and the time spent in each slide [2].

Peripheral blood smears are microscopic slides arranged from a blood sample that agree to microscopical inspection of blood cells. For inspecting blood cells, blood smears are stained. Giemsa stained blood smears are utilized to distinguish platelets, white blood cells, red blood cells and parasites [3]. Giemsa staining solution always stains up nucleic acids such that white blood cells, platelets, and parasites are colored in a dark purple color. And red blood cells are frequently painted in slight pink colors.

Automatic recognition of blood cells in images of Giemsa stained blood smears usually consists of four major phases: preprocessing, cell segmentation, feature extraction and recognition. Cell segmentation phase is always performed before cell recognition phase to extract the sharps and areas of cells from images of Giemsa stained blood smears. The segmentation phase is one of the most important key in whole recognition system because the accuracy of the subsequent feature extraction and recognition phases depend on the correct segmentation of blood cells. It is a difficult and challenging task due to the Giemsa stained blood smears have the
following complexity and uncertainty characteristics [4]: (i) Images are different sizes; (ii) Images exhibit high variation in color tone, illumination, and contrast; (iii) Images may contain noise; (iv) The shapes and appearances of cells may vary significantly among the images. Figure 1 shows several samples of available stained blood smear images with different colors, scalars, resolutions, noise, and illuminations.

![Figure 1. Samples of stained blood smear images.](image)

Common segmentation schemes on blood images in previous works include Canny edge detection [5], Sobel edge detection [6], gradient vector flow for snake [7], watersheds [8], k-means clustering [9], fuzzy c-means clustering [10], support vector machine [11], and artificial neural network [12], etc. For cell segmentation, Mat-Isa et al. proposed a seeded region growing feature extraction (SRGFE) scheme to obtain the 4 features of a certain region of interest (ROI) on a digital cancer cell image. In their scheme, the user needs to determine the ROI by clicking the mouse on any pixels in the ROI and to give the thresholding value [13]. Jiang et al. presented a white blood cell (WBC) segmentation algorithm based on scale-space filtering and watershed clustering techniques. In their WBC scheme, nucleus and cytoplasm are extracted with two different methods, respectively; a sub image containing WBC is first separated from the original cell image, nucleus region is then extracted with employing a scale-space filtering on sub image, cytoplasm region is successively extracted by processing a watershed clustering in histogram, the entire connective WBC region is finally extracted by performing morphological operations [14]. Guo et al. directly applied Support Vector Machine (SVM) on the spectrum of each pixel for white blood cell segmentation, and utilized sequential minimal optimization algorithm to select features to reduce the training time of SVM classifier. Experimental results showed that their scheme is effective, insensitive, and robust to smear staining and illumination conditions [15]. The proposed multiple cells segmentation algorithm is composed of the color spaces transformation scheme, image histogram equalization, Outliers detection, Otsu thresholding scheme, modified significance linked connected component analysis (MSLCCA) algorithm, Canny edge detection, mathematic morphological open operation and cell size filtering method. The color spaces transformation scheme transforms color images of Giemsa stained blood smears into gray scale image to simplify the segmentation algorithm to save operation time. The noise in the gray scale image is then filtered out by the median filter. For a noise filtered image of blood smears, the proposed algorithm first detects the lower outliers in the histogram of the gray scale
image to extract white blood cells and platelets. Otsu thresholding scheme is employed to binarize the white blood cells and platelets removed image. Candidate red blood cells in the binary image are labeled and extracted by the MSLCCA algorithm. Canny edge detection is conducted on each extracted red blood to evaluate its corresponding average radius and maximum diameter to select red blood cell with suitable size. The candidate red blood cells whose sizes are not big enough are removed as noises. On the other hand, candidate red blood cells whose sizes are not small enough are selected as overlapped red blood cells; the overlapped red blood cells are separated by mathematic morphological operations. Mathematic morphological operations scheme is repeated with increasing the sizes of its kernel elements until there is not any overlapped red blood cell in the image. The details of the proposed algorithm are illustrated in section 2. The experimental results are presented in section 3. Finally, section 4 concludes this paper.

2. The proposed algorithm

The main goal of the proposed algorithm is to extract the white blood and red blood cells from images of Giemsa stained blood smears for doctors to efficiently and accurately diagnose diseases. The flow chart of the proposed algorithm is shown in Figure 2. The proposed algorithm consists of color spaces transformation scheme, image histogram equalization, Outliers detection, Otsu thresholding scheme, modified significance linked connected component analysis (MSLCCA) algorithm, Canny edge detection, mathematic morphological operations and cell size filtering method. The detail blood cell segmentation steps are illustrated in following subsections.

2.1 Color mapping RGB to YIQ.

The image in RGB color space is not suitable for image processing applications, because the image in RGB color space is highly correlated. Other color models like as HIS, L*a*b*, YIQ, YUV, and YCbCr are suitable for image processing applications, they are the reducing redundancy models of the image in RGB color space, obtained by
some color transform. A common alternation to the RGB representation of an image is the YIQ representation. The YIQ representation of an image is the standard model in the television transmission. The YIQ representation of an image obtained from the RGB representation of an image is given by equation (1).

\[
\begin{bmatrix}
Y \\
I \\
Q
\end{bmatrix}
= \begin{bmatrix}
0.299 & 0.587 & 0.144 \\
0.596 & -0.274 & -0.322 \\
0.212 & -0.523 & 0.311
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\] (1)

\[
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
= \begin{bmatrix}
1.000 & 0.956 & 0.621 \\
1.000 & -0.272 & -0.647 \\
1.000 & -1.106 & 1.703
\end{bmatrix}
\begin{bmatrix}
Y \\
I \\
Q
\end{bmatrix}
\] (2)

2.2 Extractions of white blood cells and platelets using outliers detecting

In statistics, an outlier is defined as an observation that deviates substantially from other observations as to be considered that it was generated by a different system. Outliers frequently have an effect on the parameters estimating of a model that being fitted to the data. This could cause inaccurate predictions and mistaken conclusions. In many cases, outliers are frequently removed to improve accuracy of the estimators. To define the outlier of a data set, let be \( \bar{x} \) the mean and let \( \sigma \) be standard deviation of the data set. One observation is declared as an lower outlier if it is no more than \( (\bar{x} - k\sigma) \), declared as an upper outlier if it is no less than \( (\bar{x} + k\sigma) \), and the others are declared as inliers, where the value of \( k \) is usually taken as no more than 2 and no less than 3.

While a staining cell for observation under a microscope, a white blood cell is always darker than a red blood cell because the white blood cell has a nucleus. Moreover, a platelet is smaller than a red blood cell, and is also darker than a red blood cell. The darkest areas in a noise removed gray-level blood image are white blood cells and platelet regions. These dark regions are always located in the lower outliers of the pixel value distribution of a gray-level blood image. It should be a good approach that to localize the white blood cells and platelets by detecting the lower outliers of the histogram of a gray-level blood image. For white blood cell and platelet detection, the proposed algorithm calculates the mean and standard deviation of a gray-level blood image to take the points of lower outliers of the blood image as the candidate white blood cell and platelet points.

2.3 Region merging with modified SLCCA

Chai et al. [16] discovered that the significant coefficients in wavelet field are only loosely clustered, the conventional definition of connected component will produce too many components and affecting the coding efficiency. So they used symmetric structuring elements with a size large than a 3x3 square to connect components that are not geometrically connected to enhance the coding efficiency. The progressive cluster detection using conditioned dilation operation is illustrated in Figure 3, where the image size is assumed as 7x7 and 8-connected structuring element is used. Moreover, white pixels are insignificant coefficients that are not coded, black pixels denote encoded significant (S) coefficients, dark-gray pixels denote that have been checked before and encoded insignificant (I) coefficients, and light-gray pixels are new 8-connected structuring elements and will be checked consequentially. The explanatory example about significant map is shown in Figure 2(a). Where, the order of examining is up, then down, then left, then right, then left-up, then right-up, then left-down, and finally right-down under 8-connectivity structure. The pixel at (2, 2) is chosen as a seed, and the remaining steps of the recursive cluster detection are shown in Figure 2 (b)-(p).

In this paper, the SLCCA is hired to construct a fast and effective region
segmentation scheme. However, we modify the representation of the SLCCA as follow:
(a) if the neighbor pixel is similar to the seed pixel, then they are in the same cluster denoted “same” (S), otherwise denoted “different” (D), and N represents that it has been processed already. (b) Each seed pixel has 8 neighboring pixels, therefore has 8 values S, D or N regularly. (c) According to (a) and (b) we start from a seed to examine the seed and record its neighboring pixels, and then repeat the same steps to examine all pixels of an image, finally we connect these (S) that are grouped together continuously to form a region. By modified SLCCA, the example of Figure 3, its final transmitted linking code string representation that a cluster included is DDDSDDDS NNNDDNND NSNNNDS NDNSNSSS DNNNSNSND DNNDDDDN NNNNNNND NSNSDDSDS NSDNNNSN NDNNDNDD NNDNDNDN.

Figure 3. Demonstration of the progressive SLCCA: (a) The significance map, (b) the seed and its 8-connectivity, and (c)-(p) steps of the algorithm.

2.4 Binarization using Otsu’s algorithm[17]

The binarization step is used to obtain a red blood cell rough contour from an eye image. On the other hand, the gray value of the real border of the red blood cell is frequently similar to that of the background in a grayscale blood image. So, the presented algorithm adapts the famous Otsu thresholding scheme to binarize a grayscale blood image to determine the red blood cell rough contour. The Otsu thresholding scheme proposed by Otsu (1979) searches an optimal threshold to divide a grayscale image’s pixels into two classes. The optimal threshold is evaluated by the discriminated criterion which maximizes the separability between target and background classes. The Otsu thresholding scheme inputs a data set and determines the maximum and minimum values of the input data set, indicated as $L_{\text{min}}$ and $L_{\text{max}}$, respectively. The histogram of the data set is normalized as a probability distribution by the following equation.

$$p(l) = n(l)/N, \quad p(l) \geq 0, \int_{l_{\text{min}}}^{l_{\text{max}}} p(l) \cdot dl = 1$$ (3)

Here, $n(l)$ is the number of elements with value $l$ and $N$ is the total number of elements of the data set. We suppose that all elements of the set are divided into two classes, $C_1$ and $C_2$ by a threshold $k$. Then, the probabilities of occurrence, $\omega$, and mean $\mu$, of each class are evaluated by the following formulas.

$$\omega_1 = \Pr(C_1) = \frac{1}{N} \int_{L_{\text{min}}}^{k} n(l) \cdot dl = \omega(k)$$ (4)

$$\omega_2 = \Pr(C_2) = \frac{1}{N} \int_{k}^{L_{\text{max}}} n(l) \cdot dl = 1 - \omega_1 = 1 - \omega(k)$$ (5)
\[ \mu = \int_{-\infty}^{\mu} l^* \Pr(l|C) \, dl = \int_{-\infty}^{\mu} l^* p(l|\omega_1) \, dl = \mu(k) / \omega_1(k) \]  \hspace{1cm} (6) \\
\[ \mu_2 = \int_{-\infty}^{\mu} l^* \Pr(l|C) \, dl = \int_{-\infty}^{\mu} l^* p(l|\omega_2) \, dl = \frac{\mu - \mu(k)}{1 - \omega(k)} \]  \hspace{1cm} (7) \\

Where

\[ \mu = \mu(L_{\text{max}}) = \int_{-\infty}^{L_{\text{max}}} l^* p(l) \, dl \]  \hspace{1cm} (8)

The class variances are evaluated by

\[ \sigma_w^2 = \int_{-\infty}^{\mu^*} (l - \mu)^2 \Pr(l|C) \, dl = \int_{-\infty}^{\mu^*} (l - \mu)^2 p(l|\omega_1) \, dl \]  \hspace{1cm} (9) \\
\[ \sigma_b^2 = \int_{-\infty}^{\mu^*} (l - \mu)^2 \Pr(l|C) \, dl = \int_{-\infty}^{\mu^*} (l - \mu)^2 p(l|\omega_2) \, dl \]  \hspace{1cm} (10)

The within-class variance, the between-class variance and the total variance of element-values are defined as follows.

\[ \sigma_w^2 = \omega_1 \sigma_1^2 + \omega_2 \sigma_2^2, \]  \hspace{1cm} (11) \\
\[ \sigma_b^2 = \omega_1 (\mu_1 - \mu) \^2 + \omega_2 (\mu_2 - \mu) \^2 = \omega_1 \sigma_1 + \omega_2 \sigma_2 \]  \hspace{1cm} (12) \\
\[ \sigma_t^2 = \omega_1 \sigma_1^2 + \omega_2 \sigma_2^2 + \int_{-\infty}^{\mu^*} (l - \mu)^2 p(l) \, dl \]  \hspace{1cm} (13)

Otsu introduced the following two measuring functions to get the optimal threshold \( \hat{k} \), they are shown as follows:

\[ \hat{k} = \arg \max_{l_{\text{min}} \leq k \leq l_{\text{max}}} (\sigma_b^2(k)) = \arg \max_{l_{\text{min}} \leq k \leq l_{\text{max}}} \left( \omega_1 \sigma_1^2 + \omega_2 \sigma_2^2 \right) \]  \hspace{1cm} (14)

In general, the bigger black regions in a binary blood-image obtained by Otsu filter will contain the red blood cell region and other smaller black regions may be the noises. So, the proposed algorithm labels these black regions and evaluates their size to extract the bigger black regions as candidate red blood cell regions.

### 2.5 Edges detection

Edges form the outline of an object and characterize the boundaries between different objects. If the edges in an image can be detected accurately, all of the objects in the image can be localized and the basic properties of objects such as size, area and shape can be defined. The edges for an image are frequently considered as important characteristics that can be extracted by applying detectors. Edge detection filters out useless data and preserves the important information in an image; it is helpful for image segmentation. In this paper, we adopt Gaussian filter and Canny edge detector to detect the outer boundaries of red blood cells.

Gaussian filter: To reduce the noise presented in images is an important aspect of image processing: Noise mixed to an image always degrades the image to such a result that important features are no longer observable. Spatial filtering of images is an important aspect of image processing as it provides a way for filtering out noise and sharpening blurred images. There are many types of spatial filtering operations: the median filter is a simple linear spatial filter. It is good at reducing Gaussian noise present in an image, but in the presence of salt-and-pepper noise, it will not perform well. Another popular linear spatial filter is the Gaussian filter [18]. This filter is circularly symmetric with giving a higher weighting values to where are closer to the
center pixel. It is also a lowpass smoothing filter which is one of the extremely successful filters in computer vision for reducing noise, calculating the gradient intensity change of an image. The Gaussian filter is simply a moving-average mask constructed with a set of coefficients. The coefficients $C(i, j)$ of the Gaussian matrix (mask) are the normalized values of the following Gaussian function $G(i, j)$ over the entire filter window.

$$G_{u, v}(i, j) = \exp\left\{-\frac{(i-u)^2 + (j-v)^2}{2\sigma^2}\right\}$$  

(16)

where $(i, j)$ is the coordinate of a pixel within the filter window, $(u, v)$ is the coordinate of the centre of the filter window, and $\sigma^2$ is the variance of the Gaussian distribution $G$. The $\sigma$ value enables the user to adjust the amount of spatial averaging that occurs in the image. The grey level $F(u, v)$ of a filtered pixel located at $(u, v)$ with a Gaussian filter is the sum of $C(i, j) \times P(i, j)$ over all pixels in the filter window, where $P(i, j)$ is the original value at location $(i, j)$.

$$F(u, v) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} P(i, j) \times C(i, j)$$  

(17)

Where $F(u, v)$ is the filtered value of the target pixel with coordinate $(u, v)$, $n$ is the vertical/horizontal dimensions of the Gaussian matrix(mask). The dimension $n$ of a Gaussian mask enables the user to adjust the blurring of the edges and number of fine lines in an image. We apply a Gaussian smooth filter to soften edges, and to reduce spurious points (noise) in an image before performing object extracting.

The Canny Edge Detection: Edge detection is the process of locating the edge pixels; it is also a pre-processing step in almost all image processing since it can simplify the analysis of images by significantly reducing the amount of data to be processed. It is applied in automatic inspection, image analysis and organ model reconstruction.

Many algorithms have been proposed for the edge detection of images. The Canny edge detector is one of the most used in computational image processing [19]. In 1986, John Canny defined an ideal edge detector which optimizes the three criteria:

(i). Good Detection: the edge detector should respond only to existing edges, and should detect all of them. This means that the detector should maximize the signal-to-noise ratio ($SNR$).

(ii). Good Localization: the points belonging to a detected edge should be as close as possible to the true edge. This means that the detector should maximize the localization; Localization ($L$) is defined as the reciprocal of the position variance of the detector response.

(iii). Low Multiple Response (LMR): the detector should minimize the number of response to a single edge. Canny used the variation approach and the boundary conditions to find the detector $g$ that maximizes the product $SNR \times L$ subject to the Low Multiple Response constraint. Then he proposed the first derivative of Gaussian function as an approximation to his detector. In two dimensions, the Canny edge detector $g_n(x, y)$ is the gradient of the two dimensional Gaussian distribution $G(x, y)$ with standard deviation $\sigma$.

### 3. Experimental results

In this paper, the experiment uses three RGB color images of Giemsa stained blood smears taken from internet with different sizes, colors, and illuminations. Figure 4 shows an example of processing result of the
Figure 4. An example of processing result of the proposed algorithm on an image of Giemsa stained blood smear.
Figure 5. The statistical analysis of each stage of the proposed algorithm.

Figure 6. The extraction results of the proposed algorithm.
Figure 7. Boxplots of the statistical analysis of the extraction results; (a) radii of red blood cells, (b) areas of red blood cells, (c) areas of double-concave dishes, (d) area ratio of double-concave dish over red blood cell.

4. Conclusion

Many diseases such as Cancer, Huntington’s disease, Malaria, Multiple Sclerosis, and Myalgic Encephalomyelitis are mainly diagnosed by the appearances include size, shape, color and the double-concave dish size of red blood cells in the blood smear. The automatically evaluating process in the appearances from the blood smear images is high importance for rapid and accurate diseases diagnosis. The proposed multiple cells segmentation algorithm is composed of the color spaces transformation scheme, image histogram equalization, Outliers detection, Otsu thresholding scheme, modified significance linked connected component analysis (MSLCCA) algorithm, Canny edge detection, mathematic morphological open operation and cell size filtering method to achieve a fully automated detection and segmentation of blood cells from the background and isolating, overlapping, or occluded cell. The color spaces transformation scheme transforms color images of Giemsa stained blood smears into gray scale image to simplify the segmentation algorithm to save operation time. The noise in the gray scale image is then filtered out by the median filter. For a noise filtered image of blood smears, the proposed algorithm first detects the lower outliers in the histogram of the gray scale image to extract white blood cells and platelets. Otsu thresholding scheme is
employed to binarize the white blood cells and platelets removed image. Candidate red blood cells in the binary image are labeled and extracted by the MSLCCA algorithm. Canny edge detection is conducted on each extracted red blood to evaluate its corresponding average radius and maximum diameter to select red blood cell with suitable size. The candidate red blood cells whose sizes are not big enough are removed as noises. On the other hand, candidate red blood cells whose sizes are not small enough are selected as overlapped red blood cells; the overlapped red blood cells are separated by mathematic morphological operations. The iterative separated red blood cell filtering (SRBCF) stage can separate cell compounds iteratively by different radii disks to enhance the performance of the cell compounds separating. The experimental results show that the proposed algorithm has precise segmentation ability even when the blood image is in poor quality and complex background. The proposed scheme can further be developed to an image recognition system for cells to provide an efficient and precise examination and diagnosis for doctors. In the future, other new schemes could be employed to strength the accuracy and integrate other image recognition methods to construct a complete blood cell recognition system.

References


