Automatic recognition of five types of white blood cells in peripheral blood

Seyed Hamid Rezatofighi\textsuperscript{a,1}, Hamid Soltanian-Zadeh\textsuperscript{a,b,*}

\textsuperscript{a} Control and Intelligent Processing Center of Excellence, Electrical and Computer Engineering Department, University of Tehran, Tehran 14395-515, Iran
\textsuperscript{b} Image Analysis Laboratory, Radiology Department, Henry Ford Health System, Detroit, MI 48202, USA

\textbf{A R T I C L E   I N F O}

\begin{tabular}{l}
Article history: \\
Received 7 September 2010 \\
Received in revised form 1 January 2011 \\
Accepted 10 January 2011 \\

Keywords: \\
White blood cell \\
Peripheral blood \\
Gram–Schmidt orthogonalization \\
Segmentation \\
Texture feature \\
Feature selection \\
Classification \\
\end{tabular}

\textbf{A B S T R A C T}

This paper proposes image processing algorithms to recognize five types of white blood cells in peripheral blood automatically. First, a method based on Gram–Schmidt orthogonalization is proposed along with a snake algorithm to segment nucleus and cytoplasm of the cells. Then, a variety of features are extracted from the segmented regions. Next, most discriminative features are selected using a Sequential Forward Selection (SFS) algorithm and performances of two classifiers, Artificial Neural Network (ANN) and Support Vector Machine (SVM), are compared. The results demonstrate that the proposed methods are accurate and sufficiently fast to be used in hematological laboratories.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Recognition and inspection of white blood cells in peripheral blood can assist hematologists in diagnosing diseases like AIDS, leukemia, and blood cancer, making it one of the most salient steps in hematological procedures. This analysis can be accomplished by automatic and manual approaches. Automatic methods usually involve instruments such as flow cytometry and automatic counting machines. These instruments can examine white blood cells quantitatively but not qualitatively; they do not benefit from image processing techniques. Applying automatic systems that include image processing techniques may provide qualitative evaluation and thus enhanced judgments. Furthermore, some of these tasks such as manually scrutinizing blood cells by experts are tedious and susceptible to error. Therefore, an automatic system based on image processing techniques can help the hematologists and expedite the process.

Although not extensive, some methods are proposed in the literature for this purpose. Generally, in these articles, preprocessing, segmentation, feature extraction, and classification are introduced as four major steps of analyzing hematological images. We describe these steps, one by one, in the following.

The main purpose of preprocessing is to generate low noise, high contrast images for the next steps. For example, in [1], a median filter is applied for noise reduction and histogram equalization is used for contrast enhancement. In [2], spatial filtering along with morphological operators are used to eliminate black and white stains.

Segmentation is the most challenging step and thus improvement of nucleus and cytoplasm segmentation is the most widespread effort in many researches. In [1,3–6], the authors suggested several methods to segment nuclei of white blood cells via techniques that can be categorized into color-based methods. These methods are simple but are not capable of segmenting the white blood cells nuclei accurately. In addition, cytoplasm is colorless in most cases. Thus, its boundary is not detectable and cannot be segmented by these methods. Methods based on imaging techniques generate superior results. For example, the method proposed in [7] obtained acceptable results using multi-spectral imaging techniques. In this method, intensity of each pixel in different spectra is used to construct the feature vectors and a Support Vector Machine (SVM) is used for classification and segmentation. In spite of efficacy of this method for segmenting white blood cells components, its implementation is costly and thus cannot be used in all laboratories. Cytoplasm and nucleus segmentation via mathematical and contour models is the third method and also the most important one. In this category, region growing [8], watershed [9], parametric active contour deformable models [10], and combination of the watershed technique and a parametric deformable model [2] are introduced in the literature. These methods are more complex and require more processing time compared to other methods. However, they have the advantage of being more accurate. On the other
hand, the initial contours selected in parametric deformable models affect the final results. In [10], the initial contour is obtained in 3 steps. First, nucleus region is segmented by image thresholding. Next, morphological operations like opening and dilation are applied. At the end, the contour of the resulting area is used as the initial contour. In this case, if white blood cells have several separate nuclei that are far from each other, this method has a limitation and generates several initial contours instead of one.

Feature extraction is usually carried out based on different types of features. Since morphological and textural features are the features extracted from white blood cells by a hematologist, many papers such as [1,11–13] use feature extraction methods on the basis of these features. Moreover, extracting textural features by the co-occurrence matrix are presented by Sabino et al. [1]. For classification, the Bayes classifier [6] and different types of Artificial Neural Networks (ANNs) such as feed-forward back propagation [14,15], local linear map [16], and fuzzy cellular neural network [17] are used in the literature.

In this paper, our main purpose is to develop a new algorithm based on image processing methods to classify five major types of white blood cells in the peripheral blood. These five groups are eosinophils, basophils, monocytes, lymphocytes, and neutrophils. Peripheral blood analysis is one of the most fundamental approaches in hematological laboratories and provides asset information for disease diagnosis. At first, nuclei of the white blood cells are segmented using a Gram–Schmidt orthogonalization method. Since in most cases, there is no clear boundary between nucleus and cytoplasm of basophils, these cells should not be involved in the segmentation of the cytoplasm. Therefore, recognizing basophils and distinguishing them from the other samples are performed using features extracted from the nucleus areas. In order to classify the remaining groups of the white blood cells, features should be extracted from the cytoplasm and nucleus areas separately. Therefore, cytoplasm segmentation is imperative. As cytoplasm edge is colorless and unobservable, using methods like deformable models and snakes are lucrative. To this end, a snake algorithm is used after the preprocessing procedures. Three kinds of features, color, morphological, and textural features, are elicited from the nucleus and cytoplasm areas. Two groups of texture features attained by the local binary pattern (LBP) and the co-occurrence matrix are evaluated. The feature selection step is adjoining to this process for improving the classifier performance and expediting the program process. Finally, the performance of two different classifiers, SVM and ANN, when using different sets of features is compared.

The rest of the paper is organized as follows. In Section 2, we introduce the methods used for automatic recognition of five types of white blood cells in hematological images. The system architecture and application of these methods are detailed in Section 3. Experimental results are presented and discussed in Section 4. Finally, Section 5 presents the conclusions of the work.

2. Methods

2.1. Segmentation

Segmentation is the most crucial step of hematological images analysis. An accurate segmentation allows accurate results in subsequent stages. For this purpose, our proposed method includes two levels in nucleus and cytoplasm segmentation using two techniques, Gram–Schmidt orthogonalization and snakes algorithm.

2.1.1. Gram–Schmidt orthogonalization method

The Gram–Schmidt procedure orthogonalizes a set of vectors in an inner product space, most commonly the Euclidean space $\mathbb{R}^n$.

It translates a finite, linearly independent set $S = \{v_1, \ldots, v_n\}$ to an orthogonal set that span the same subspace. To depict the relations in this algorithm, a projection operator is defined as follows:

$$
\text{proj}_{u}^{v} = \frac{\langle u, v \rangle}{\langle u, u \rangle} u
$$

(1)

Here $\langle u, v \rangle$ denotes the inner product of vectors $u$ and $v$. This operator projects $v$ onto the vector $u$. According to this definition, the Gram–Schmidt orthogonalization procedure is as follows:

$$
\begin{align*}
   u_1 &= v_1, \\
   e_1 &= \frac{u_1}{\|u_1\|}, \\
   u_2 &= v_2 - \text{proj}_{u_1}^{v_2}, \\
   e_2 &= \frac{u_2}{\|u_2\|}, \\
   u_3 &= v_3 - \text{proj}_{u_1}^{v_3} - \text{proj}_{u_2}^{v_3}, \\
   e_3 &= \frac{u_3}{\|u_3\|}, \\
   & \vdots \\
   u_k &= v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j}^{v_k}, \\
   e_k &= \frac{u_k}{\|u_k\|}.
\end{align*}
$$

(2)

The sequence $u_1, \ldots, u_k$, is the required system of orthogonal vectors and the normalized vectors $e_1, \ldots, e_k$ form an orthonormal set [18].

Using this method, a vector $w_k$ can be attained for the set of vectors $S = \{v_1, \ldots, v_n\}$ such that it has maximal projection on to $v_k$ and is orthogonal to the other vectors in the set:

$$
\begin{align*}
   w_k &= v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j}^{v_k}, \\
   \langle v_k, w_k \rangle &= K, \ K \neq 0.
\end{align*}
$$

(3)

Therefore, the inner product of the set $S$ in $w_k$ is:

$$
\begin{align*}
   \langle v_j, w_k \rangle &= 0, \ j \in 1, \ldots, n \text{ and } j \neq k, \\
   \langle v_k, w_k \rangle &= K, \ K \neq 0.
\end{align*}
$$

(4)

Fig. 1 illustrates the relationship between $w_k$ and vectors $v_1, v_2$, and $v_3$ in the 3D space.

In order to use a Gram–Schmidt process for segmentation of color images, each feature vector is considered as a vector in this method. Elements of this vector can be pixel intensities in the RGB space. Calculating $w_k$ on the basis of the Gram–Schmidt process and applying it to a color image intensify the desired color, represented by $v_k$, and attenuate the other colors. As a result, a composite image is obtained in which the regions of interest with the desired color have maximum intensity whilst other regions have minimum intensity. Hence, an appropriate threshold may generate a desirable segmentation.

2.1.2. Parametric deformable models and snakes

Parametric deformable models or snakes are energy-minimizing splines guided by the internal constraint forces and influenced by the image forces that pull them towards distinct
features (e.g., lines and edges) [19]. The snake algorithm is an iterative process used for object segmentation when edges are not well-defined. It starts from an initial contour around the object. By taking into account quantitative energy values for the points that make up the snake, the algorithm moves the positions of the snake points to minimize the overall energy of the snake. The energy function for the snake defined in [19] and represented by the vector $c(s) = (x(s), y(s))^T$ having arc length $s \in [0,1]$ as a parameter is:

$$E(c) = \int_0^1 (E_{\text{int}}(C(s)) + E_{\text{ext}}(f, C(s)))ds$$

(5)

where $f$ is the image intensity and $E_{\text{ext}}$ is the image forces. Also, $E_{\text{int}}$ represents the internal energy of the snake due to bending or discontinuities. It is defined as follows:

$$E_{\text{int}} = \int |c|^2 ds$$

$$C_t = \frac{\partial C}{\partial s}$$

$$C_s = \frac{\partial C}{\partial s}$$

$$C_{ss} = \frac{\partial^2 C}{\partial s^2}$$

$$\alpha(s) = (\alpha(s) C_t(s))^2 + \beta(s) C_{ss}(s)^2$$

Thus, in Eq. (5), the internal energy consists of $\int |c|^2 ds$ and $\int |C_s|^2 ds$ terms which relate to the curve elasticity and bending, respectively. The curve length is confined by the parameter $\alpha$. As $\alpha$ decreases, the role of the curve length in the internal energy diminishes. Furthermore, the parameter $\beta$ controls fluctuation and bending of the contour. These parameters are positive numbers and control the internal energy.

The $E_{\text{ext}}$ depends on $f$ and can be expressed as:

$$E_{\text{ext}} = W_{\text{line}} \cdot E_{\text{line}} + W_{\text{edge}} \cdot E_{\text{edge}} + W_{\text{term}} \cdot E_{\text{term}}$$

(7)

where $E_{\text{line}}$ is the simplest form of an external force and is equivalent to intensity $f(x,y)$ and the polarity and amount of $W_{\text{line}}$ indicates the direction of contour and the impact of the $E_{\text{line}}$ in the external energy. The $E_{\text{term}}$ is the force to allow the contour to detect the image edges and its impact is determined by $W_{\text{term}}$. The $E_{\text{edge}}$ is a function of image gradient that pulls the initial curve towards the image boundary and its impact is determined by $W_{\text{edge}}$. The active contour method provides a robust model to extract the desired image boundary if these parameters are set appropriately. Also, this approach is more efficient when the edges of the image are not sharp. Thus, it is a proper model to segment white blood cells cytoplasm that have invisible edges.

2.2. Feature extraction and selection

White blood cells are visually distinguishable by tangible characteristics such as nucleus and cytoplasm shape, their texture, and color. Thus, in this paper, these features are quantified.

2.2.1. Morphological features

In this part, morphological features analogous to those used by hematologist are extracted. These features include nucleus and cytoplasm area, nucleus and whole cell's perimeter, number of separated parts of nucleus, mean and variance of nucleus and cytoplasm boundaries, and the ratio between cytoplasm and nucleus areas. In addition, roundness of nucleus and the cell's body is calculated by:

$$RC = \frac{P_e^2}{4\pi A_e}$$

(8)

where $P_e$ is the perimeter and $A_e$ is the area of the region.

2.2.2. Texture features

Another kind of features commonly extracted from the nucleus and cytoplasm regions is texture features. In this research, the co-occurrence matrix and the local binary pattern (LBP) are used in order to extract texture features and are defined as follows.

2.2.2.1. Co-occurrence matrix. The co-occurrence matrix is a symmetric matrix constructed on the basis of the image gray levels with the distances $d$ and angles $\phi$. This matrix describes the second order probabilistic features [20]. By diverse $d$ and $\phi$, disparate co-occurrence matrices are created. Assuming $N_g$ as the number of gray levels in the image, the co-occurrence matrix dimension is $N_g \times N_g$. For example, the co-occurrence matrix for a simple case with $d=1$ and $\phi=0$ is derived and depicted in Fig. 2.

Fourteen features are extracted from the co-occurrence matrix to represent contrast, homogeneity, entropy, and other quantities that represent texture [20]. To make features rotation invariant, 4 matrices are usually computed at 4 angles and their average is used to calculate the 14 features. These 14 features are more robust and subtle than most of the other texture features. A long processing time is the main weakness of this method.

2.2.2.2. Local binary patterns (LBP). Local binary pattern (LBP) is a texture analysis method that analyzes texture in different radii and thus can be considered as a multi-resolution method. Two features are usually extracted from each radius. The first one is $LBP^{1 \text{st}}$, which represents the structure of texture and the other one is $VAR$ which depicts changes in gray levels [21], as described below.

(1) $LBP^{1 \text{st}}$: To extract this feature, the intensities of the central point of the image and $P$ points of the periphery with radius $R$ are extracted (Fig. 3(a)). The coordinates of the $P$ points are obtained by

$$x, y = \left( R \sin \left( \frac{2 \pi p}{P} \right), R \cos \left( \frac{2 \pi p}{P} \right) \right)$$

(9)
Next, an interpolation is applied to ascribe an intensity to each point. For determining the first feature, \( P \) points are placed in an array. In order to make the features intensity invariant, the intensity of central point \( g_c \) is subtracted from the intensities of the peripheral points.

\[
T = \{ (g_0 - g_c, g_1 - g_c, \ldots, g_{P-1} - g_c) \}
\]  

(10)

Applying the Sign function leads to a binary vector \( T_2 \):

\[
S(x) = \begin{cases} 
1 & x \geq 0 \\
0 & x < 0 
\end{cases}
\]

(11)

\[
T_2 = (S(g_0 - g_c), S(g_1 - g_c), \ldots, S(g_{P-1} - g_c))
\]

(12)

Consequently, the first feature is defined as:

\[
LBPP_{R} = \sum_{p=0}^{P-1} S(g_p - g_c) \cdot 2^p
\]

(13)

However, this feature is not rotation invariant. To obtain a rotation invariant form of the feature, the following algorithm is used. To this end, circular shifts are applied to the elements of the vector \( T_2 \) until it represents the smallest decimal value possible, as shown in Fig. 3(b). Also, a homogeneous definition may be applied to simplify the process. A function \( U \) is used to count the number of fluctuations between 0 and 1 in the binary sequence. Finally, this feature is defined by:

\[
LBPP_{R} = \left\{ \begin{array}{ll}
\sum_{p=0}^{P-1} S(g_p - g_c), & U(LBPP_{R}) \leq 2 \\
\sum_{p=0}^{P+1}, & \text{Otherwise}
\end{array} \right.
\]

(14)

(2) \( \text{VAR} \): A feature that quantifies the variation in the intensities of the \( P \) points is defined by:

\[
\text{VAR}_{R} = \frac{1}{P} \sum_{p=0}^{P-1} (g_p - \mu)^2
\]

(15)

where \( \mu \) is defined as:

\[
\mu = \frac{1}{P} \sum_{p=0}^{P-1} g_p
\]

(16)

Utilizing these two features in different radii, the tissue characteristics can be interpreted appropriately. In this paper, the feature \( \langle LBPP_{R} \rangle \), after the rotation invariant procedure is also used. Therefore, three features are totally extracted from each radius.

2.2.3. Feature selection

Feature selection is used to reduce the dimensionality of feature vectors and generate a new feature vector with a higher discrimination power. It has been shown that the classifier performance maybe improved as a result of eliminating irrelevant features. Fisher’s discriminant ratio (FDR) maybe used to select informative features. This function for \( M \) classes is as follows:

\[
FDR = \sum_{j=1}^{M} \sum_{i=1}^{M} \frac{(\mu_j - \mu_i)^2}{\sigma_i^2 - \sigma_j^2}
\]

(17)

where \( \mu_i \) and \( \sigma_i^2 \) are vectors attained by the average and variance of the features belonging to the i-th class, respectively. The higher the FDR, the better discrimination between the two classes. To select appropriate features, a specific probing algorithm is needed. In this paper, Sequential Forward Selection (SFS) method is used as the probing algorithm.

2.2.3.1. Sequential Forward Selection (SFS).

This method starts with the feature that has the maximum discrimination among the features on the basis of the FDR function. Then, all couples of features including the selected feature are given to the discriminatory function. After that, the best couple is chosen. In the next level, the best triple vector of features is found that includes the selected couple of features. The algorithm continues until the best \( M \) features are chosen [22].

After selecting features by SFS, the features are classified as described below.

2.3. Classifiers

In this paper, Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are used to classify the feature vectors and compare the results.

2.3.1. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are computational models that simulate structure and function of biological neural networks. These networks have diverse applications in machine learning. One of their applications is in classification and decision making based on existing data. Training is an important task in utilizing the neural networks. For this purpose, the input data is often divided into two parts for training and testing. Artificial Neural Networks used often are multi-layer perceptron (MLP), recurrent networks, and radial basis function (RBF) [23].

In this paper, MLP is used to classify the features extracted from the white blood cells. The network is trained using the back-propagation algorithm. The MLP uses a non-linear function, in the hidden layers, and appropriate weights between the layers so that a particular input leads to a specific target output. The number of neurons in the last layer depends on the number of classes. The number of neurons in the first layer is as same as the dimension of the input features [23].

2.3.2. Support Vector Machine (SVM)

The SVMs are new classifiers that have shown superior performance in comparison with neural networks in some applications [24]. Assuming the learning data \( \{ (x_i, y_i) \}_{i=1}^{N} \) where \( x_i \in \mathbb{R}^m \) are the inputs and \( y_i \in \{ \pm 1 \} \) are the outputs, the SVM functions as follows.

At first, the input data is registered to the Hilbert (\( F \)) space according to \( Z = \phi(x) \) which has high dimensionality. Supposing that the data in the \( F \) space is linearly separable, vectors \( W \in F \) and a scalar \( b \in \mathbb{R} \) exist such that:

\[
y_i(W, \phi(x_i)) + b \geq 1
\]

(18)

The SVM creates a linear hyperplane in the \( F \) space that has the maximum separation. Maximizing the distance between the marginal data points of any class and the hyperplane is the most important advantage of the SVM (Fig. 4).

The weight \( W \) is derived as \( w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \) where \( \alpha_i \geq 0 \). Also, \( \alpha_i \) can be derived by the expression below and quadratic programming:

\[
\begin{align*}
L_D = & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle \phi(x_i), \phi(x_j) \rangle \\
\end{align*}
\]

(19)

In practice, the Kernel functions \( (K) \) are applied instead of the \( F \) registration with infinite dimensions as:

\[
K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle
\]

(20)
Fig. 4. Maximum distance between data points on the edge (support vectors) and boundary (linear hyperplane) in SVM.

The Mercer condition holds in this Kernel function. In this paper, the Gaussian Kernel function is utilized as:

$$K(x_i, x_j) = \exp\left(\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$  \hspace{1cm} (21)

When the data in the $F$ space is not linearly separable, another parameter is defined as a trade-off between model complexity and learning errors [25].

3. System architecture

Designing an automatic system to recognize five types of white blood cells in a hematology image of the peripheral blood is the main goal of this work. To achieve this goal, a block diagram is designed based on this type of dataset. Fig. 5 illustrates the block diagram of our proposed system. As shown in this figure, the method has three major phases:

Phase I: Nucleus segmentation, nucleus feature extraction, and classification to distinguish basophils from other kinds of white blood cells.

Phase II: Reducing the image size, detecting the initial contour for snake, and preprocessing for performing the snake algorithm.

Phase III: Cytoplasm segmentation, feature extraction from the nucleus and cytoplasm region simultaneously, classification and recognition of the remaining four groups of white blood cells.

The details of these phases are explained in the following sections.

3.1. Phase I

Since in most of the samples, the regions belonging to the nucleus and cytoplasm of basophils cannot be distinguished visually, these cells should not be involved in the cytoplasm segmentation step. Therefore, they should be distinguished from the other cells in this phase.

Segmentation of nucleus by Gram–Schmidt orthogonalization method: In this method, the pixel intensities of the RGB components for a hematology image are considered as 3D vectors. $v_1$ as a desired vector is obtained by averaging the 3D vectors of the nucleus area in some samples. $v_2$ and $v_3$ as undesired vectors are defined from the areas which are similar to the nucleus but are not from the nucleus area. Using $v_1$, $v_2$ and $v_3$, a weighting vector $w$ is attained whose inner product with the pixel vectors generates a composite image with higher intensity in the nucleus area compared to other areas (Fig. 6(a)). Because the samples are diverse in intensity and color, using Gram–Schmidt method as stated above cannot be applicable for segmentation in all classes. Thus, we use three different sets of the Gram–Schmidt method as their results are complementary. The three different sets of $v_1$, $v_2$ and $v_3$ result in three different vectors for $w$. Next, to segment the nucleus, an appropriate threshold is calculated for each resulting image which is adaptively attained based on the image histogram. It is important to note that after thresholding, some fine regions are enhanced in addition to the nucleus regions due to their similarity with the nucleus color. These diminutive regions are omitted with the morphological operations. At the end, a combination of the three results eventuates the desired image (Fig. 7). The final result of the nucleus segmentation is demonstrated in Fig. 6(b).
Feature extraction from the nucleus area and classification to distinguish basophils: As previously stated, basophils should be recognized and separated from the other types of white blood cells in this step. Therefore, several morphological features such as nucleus area and perimeter, number of separated parts of the nucleus, mean and variance of the nucleus boundaries, and roundness of nucleus are extracted from the segmented area. To extract textural features from this area, the co-occurrence matrix and the local binary pattern are applied and the results are compared. Color features are the other features extracted as a normalized vector of averaged nucleus color.

3.2. Phase II

In this phase, the main purpose is to prepare the image for the snake algorithm to segment cytoplasm. To this end, image size reduction, preprocessing, and finding of an initial contour for the snake algorithm are done as explained below.

Image size reduction: Using the nucleus area segmented in the previous phase, we find the center of each nucleus and fit an appropriate window around it to define a sub-image that contains a complete white blood cell. This process makes the segmentation process easier. In our research, a $141 \times 141$ window is used. As shown in Fig. 8, both of the segmented and original images are cropped by the same window.

Finding an initial contour for snake algorithm: To this end, morphological dilation is applied to the segmented nucleus regions. The structuring element for the dilation operation is a square with an adaptive size based on the nucleus size. After the dilation, if there is just a single region for white blood cell, the convex boundary of this region is used as the initial contour. Otherwise, a circle with an adaptive radius, calculated based on the nucleus size, and located in the center of the window is considered for this purpose. Fig. 9 shows
Fig. 9. (a) and (d) Segmented nuclei of two multi-nuclei white blood cells. (b) and (e) The images after dilation operation. (c) and (f) Initial contours.

Fig. 10. (a) A sample image. (b) The image after histogram equalization. (c) The image after extracting the saturated plate. (d) The smoothed image by a Gaussian kernel.

Preprocessing before snake algorithm: Due to high accumulation of the red blood cells, they may touch the cytoplasm of the white blood cells. Thus, the boundary between the cytoplasm and red blood cells may not be distinguished when the color image is changed into gray-scale. To solve this problem, the image is enhanced by the color histogram equalization. Next, the enhanced image is transferred into the Hue-Saturation-Intensity (HSI) space. The final image is attained by extracting the saturation plate from the HSI image. Based on this idea, we have a gray-scale image that has good discrimination between the boundaries of the cytoplasm and the red blood cells. Then, the image is smoothed by a Gaussian kernel to eliminate the cytoplasm cavities and inhomogeneities and ameliorates the image for the snake algorithm (Fig. 10).

3.3. Phase III

The main aim of this phase is to recognize the remaining four classes of the white blood cells. To this end, after preprocessing and initial contour detection, the snake algorithm is applied to segment the cytoplasm. After segmentation, textural and morphological features are extracted from both of the nucleus and the cytoplasm areas and the four types of cells are classified.

Cytoplasm segmentation using snake algorithm: The snake algorithm starts from the initial contour obtained in the previous phase and its parameters are set to \( \alpha = 2, \beta = 5, \gamma = 0.7, \rho = 0.4 \). The snake algorithm ends when no snake points move to new positions for four consecutive iterations.

Feature extraction from nucleus and cytoplasm, and classifying the remaining four classes: In this step, features are extracted from the cytoplasm area and used in combination with the features extracted from the nucleus area in phase I. These features may be categorized into three groups of morphological, texture, and color features. The morphological features are cytoplasm area and whole cell body perimeter, mean and variance of the cytoplasm boundaries, roundness of the whole cell and the ratio between the cytoplasm and nucleus areas. Texture features are also extracted from the cytoplasm area by the co-occurrence matrix and the local binary pattern and their results compared. At the end, a normalized vector of the average cytoplasm color is extracted.

4. Experimental results

4.1. Dataset

Samples were taken from peripheral blood of 8 normal subjects and 400 samples were obtained from 100 microscope slides. The microscope slides were smeared and stained by Gismo-Right technique and images were acquired by a light microscope (Microscope-Axioskope 40) from the stained peripheral blood using an achromatic lens with a magnification of 100. Then, these images were recorded by a digital camera (Sony Model No. SSC-DC50AP) and were saved in the BMP format. The images contain 720 × 576 pixels. All of them are color images and were collected from Hematology–Oncology and BMT Research Center of Imam Khomeini hospital in Tehran, Iran. The images were classified by a hematologist into normal leukocytes: basophil, eosinophil, lymphocyte, monocyte, and neutrophil. Also, the areas related to the nucleus and cytoplasm were manually segmented by an expert.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Basophil</th>
<th>Eosinophil</th>
<th>Lymphocyte</th>
<th>Monocyte</th>
<th>Neutrophil</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>94.70%</td>
<td>90.81%</td>
<td>88.86%</td>
<td>96.70%</td>
<td>94.05%</td>
<td>93.02%</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>–</td>
<td>95.55%</td>
<td>93.05%</td>
<td>91.23%</td>
<td>97.25%</td>
<td>91.79%</td>
</tr>
<tr>
<td>Average</td>
<td>94.70%</td>
<td>93.22%</td>
<td>90.01%</td>
<td>91.23%</td>
<td>96.23%</td>
<td>93.06%</td>
</tr>
</tbody>
</table>
4.2. Segmentation results

In order to quantitatively evaluate the results of the nucleus and cytoplasm segmentations, the following similarity measure is defined.

\[ T_s = 100 \times \frac{A_{\text{program}} \cap A_{\text{expert}}}{\max(A_{\text{program}}, A_{\text{expert}})} \]  

(22)

where \( A_{\text{program}} \) is the segmented area by the algorithm and \( A_{\text{expert}} \) is the segmented area by an expert. When these two areas are the same, \( T_s \) is 100. In Table 1, the resulting measures for each kind of white blood cells and their overall segmentation are presented.

Three discussable points can be inferred from Table 1. According to this Table, the result of nucleus segmentation for the lymphocyte class is lower than the other classes. The main reason is that the color of the cytoplasm is analogous to the color of the nucleus in many of the lymphocytes samples, especially young ones (Fig. 11(a)). Therefore, the segmentation error for this type of the white blood cell is larger than the others. Due to the inherent nature of the basophiles, their cytoplasm and nuclei appear together (they are not separated) (Fig. 11(b)). Hence, we do not have values to report for the cytoplasm of the basophiles. Also, since in most of the cases, the vitreous cytoplasm of monocytes is colorless (Fig. 11(c)), even the deformable model with a congruous preprocessing is unable to find the cytoplasm boundaries precisely. Therefore, the accuracy result of the cytoplasm segmentation is worse than those of the other classes.

4.3. Classification results

Classification is performed in two sections: discriminating the basophiles from other types of the white blood cells in phase I and recognizing the remaining classes in phase III.

To appraise the result of the proposed algorithms in classifying the white blood cells, the following confusion matrix is calculated.

\[ T_F = \begin{bmatrix} T_{11} & \cdots & T_{1M} \\ \vdots & \ddots & \vdots \\ T_{M1} & \cdots & T_{MM} \end{bmatrix} \]  

(23)

where \( T_{ij} \) is the number of samples of class \( i \) that are classified as samples of class \( j \). An ideal situation occurs when the matrix is diagonal. Accuracy of class \( K \) can be defined as:

\[ AC_k = 100 \times \frac{T_{kk}}{\sum_{j=1}^{M} T_{kj}} \]  

(24)

At the end, the confusion matrix leads to another criterion named the overall accuracy defined as:

\[ AC_{\text{overall}} = 100 \times \frac{\sum_{i=1}^{M} \sum_{j=1}^{M} T_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{M} T_{ij}} \]  

(25)

Using this criterion, the performance of the classifiers are compared.

4.3.1. Experimental results for basophile classification

To compare the performance of the texture features, two groups of features, extracted from the nucleus area, are created. These two groups include the same morphological and color features but different in texture features. Ultimately, the first group (24 features) includes texture features extracted by local binary pattern and the second group (23 features) consists of features obtained by the co-occurrence matrix. To classify these features, at first, feature dimension is reduced by a SFS algorithm and the results for MLP and SVM are compared by means of the overall accuracy measure.

Fig. 12. The overall accuracy results for the ANN and SVM using (a) local binary pattern and (b) co-occurrence matrix as texture features to distinguish basophiles.
The MLP consists of 3 layers where the hidden layer includes 40 neurons. In SVM classification, a Gaussian function is used as kernel and features are normalized between 0 and 1. The standard deviation of the Gaussian kernel parameter is 1. Fig. 12(a) and (b) illustrates the results related to the local binary pattern and the co-occurrence matrix, respectively.

Some points are construed from these figures. The first point is that reduction in dimension of features aggravates the classification as expected. The second one is that the overall accuracy does not have considerable escalation after 15 features for both ANN and SVM. This occurs, because after selecting 15 features, the differentiation between the features of each class is at a maximum and increasing the features perplexes the classifier. The third point is that the ANN classifier has more fluctuation in the overall accuracy in comparison with the SVM. These changes are because the MLP is not trained well as a result of trapping in a local minimum. In fact, if two analogous neural networks are trained with the same feature matrices, their results are not the same. The other point is that the ANN and SVM classifiers have similar performances in most of the feature dimensions, but due to the stability of the SVM in training, it may be preferred. The last point is that the features attained from the co-occurrence matrices have superior performance in comparison with the features obtained from LBP. However, the time required for calculating the first group of features is significantly higher than the second one.

Based on the above points, we selected 15 features and utilized SVM to generate an optimal performance in terms of classification accuracy and speed. The results are compared in Tables 2 and 3 where confusion matrices, accuracy, and overall accuracy are listed for the two groups. These tables illustrate that the results of classification with the features of the co-occurrence matrix are superior to those of the LBP. However, considering the trade-off between accuracy and processing time, LBP may be preferred. The ratio between the times required for feature extraction using the co-occurrence and LBP method is 20 to 1.

### 4.3.2. Experimental results for classifying the remaining four groups of white blood cells

Similar to the previous section, the performance of the two groups of features whose texture features are extracted by the LBP and the co-occurrence matrix are compared whilst these features are obtained from both of the nucleus and cytoplasm areas. The dimension of the first group which uses LBP is 33 and the dimension of the second group which uses the co-occurrence matrix is 32. For a second time, the SFS algorithm is applied to select the best features in a pre-specified dimension and the results are compared for ANN and SVM by the accuracy criterion. Fig. 13(a) delineates the overall accuracy of the SVM and ANN for the first group of features and Fig. 13(b) demarcates the result the other group of features.

According to the above figures, some points related to the performance of the classifiers are analogous to those discussed in the previous section. For instance, a reduction of the dimension of features exacerbates the classifier performance. Furthermore, the oscillation of the overall accuracy of the ANN classifier is considerably more than that of the SVM. Another point inferred from the figures is that after 10 features, curves do not have significant escalation for both of the ANN and SVM. The SVM classifier has a superior performance in this case in comparison with the ANN. In addition, the features obtained from the co-occurrence matrices generally have a superior performance.

Considering both classification accuracy and processing time, SVM classifiers and feature dimension of 10 can be considered optimal. Table 4 illustrates the confusion matrix, accuracy, and overall accuracy when 10 features including those of LBP are selected and SVM is used as a classifier.

### Table 4

<table>
<thead>
<tr>
<th>Recognized basophil</th>
<th>Recognized non-basophil</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophil</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>Non-basophil</td>
<td>5</td>
<td>168</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Table 5, the confusion matrix, accuracy, and overall accuracy with the co-occurrence matrix are demonstrated when 10 features are selected and SVM is used as a classifier. According to these tables, the classification accuracy for the two groups of features

---

**Table 2**

Confusion matrix, accuracy, and overall accuracy for 15 LBP features and SVM classifier.

<table>
<thead>
<tr>
<th>Recognized</th>
<th>Recognized non-basophil</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophil</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>Non-basophil</td>
<td>23</td>
<td>150</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3**

Confusion matrix, accuracy and overall accuracy for 15 co-occurrence features, and SVM classifier.

<table>
<thead>
<tr>
<th>Recognized</th>
<th>Recognized non-basophil</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophil</td>
<td>49</td>
<td>1</td>
</tr>
<tr>
<td>Non-basophil</td>
<td>5</td>
<td>168</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
is equal. However, noting the fact that calculation of the features extracted by the co-occurrence matrix is noticeably more computational, using LBP for the texture features is proposed for this phase as well.

5. Conclusion

This paper presented an image analysis system to recognize five groups of white blood cells in the peripheral blood. Compared to previous work in [1,4–6], we used the Gram–Schmidt orthogonalization method for the segmentation of the nucleus which can be categorized as the color-based method. Whilst applying three sets of Gram–Schmidt orthogonalizations was simple and had a short processing time, it reduced the susceptibility of the proposed method to color and intensity variations. Another novelty of the method was to propose an adaptive algorithm for finding an initial contour for the snake algorithm. This method solved a weakness of the algorithm introduced in [10] (generation of several initial contours for multi-nuclei cells).

Regarding the texture features, methods based on the LBP and the co-occurrence matrices were compared. Although they generated almost the same classification results, calculation of the features extracted from the LBP was less computational. Therefore, selection of these features for a hematological system seems attractive. Furthermore, comparison of ANN and SVM classifiers confirmed that SVM was superior and thus, the proposed algorithm used SVM.

The proposed method has a reasonable processing time and is accurate. Its overall segmentation accuracy is 93% whilst its classification accuracies in phases I and III are 90% and 96%, respectively. Regarding the processing time, the program requires 10 s for analyzing a white blood cell on a Pentium-4 PC, running at 3.2 GHz, with 1 GB of RAM using MATLAB. Hence, differential counting of 100 white blood cells lasts about 16 min. In comparison, an expert requires about 15 min to carry out this process. Thus, this program can be used in hematological laboratories. The other outputs of this program, in addition to differential counting, can be statistical scrutiny of shape, size, and the ratio between nucleus and cytoplasm of the counted cells which provide asset information for hematologists in disease diagnosis and evaluation of disease progression.

Notwithstanding mentioned advantages, the proposed method may need calibration when new datasets with different characteristics are introduced to the system. This is because in the Gram–Schmidt method, the operator defines vectors $v_1$, $v_2$, and $v_3$ for each set as an initial calibration. Although the data used in this work included a wide range of colors and intensity variations and the proposed method was not susceptible to these variations, for maximum accuracy, a calibration is recommended when using a new dataset.

As future work, an algorithm can be designed to define the three vectors in the Gram–Schmidt method automatically. In addition, it is cogent to add a new class for the white blood cells that do not belong to the five classes considered in this work. This is due to the fact that sometimes other cells called blast appear in the peripheral blood. This cell type is more frequently found in the abnormal blood samples. Moreover, adding a new step can help in appraising the characteristics of each group of white blood cells in order to inspect if cells are normal or suspected of leukemia.

Acknowledgments

The authors would like to thank Dr. Ramezanali Sharifian, Dr. Reza Safaee, and Ms. Hajir Mirmongereh for their assistance with the gathering of the hematology images. The suggestions and comments of Ms. Kosar Khaksari helped improve the quality of the paper and are appreciatively acknowledged.

References


