Automatic leukocytes classification by distance transform, moment invariant, morphological features, gray level co-occurrence matrices and SVM

Gai Pang [a], Yangkai Zhuang [b], Ping Zhou [c]

School of Information, Zhejiang Sci-Tech University, Hangzhou, 310018, China

[a]470040845@qq.com, [b]281901879@qq.com, [c]zp@zstu.edu.cn

Keywords: Leukocytes classification; Euclidean distance transform; Moment invariant; Morphological features; Texture features; Support vector machine.

Abstract. Leukocyte is an important part of the immune system. According to the problem that manual operation is not efficient, a novel automatic classification of leukocytes is proposed in this paper. First, moment invariant based on Euclidean distance transform is extracted from nucleus area and morphological features are extracted from segmented cells. Then, monocytes, lymphocytes and basophils are distinguishing from the other samples using features extracted from the first step. Next, the gray level co-occurrence matrices (GLCM) are used as texture measure to classify the remaining classes via a Support Vector Machine (SVM). Experimental results show that the proposed approach provided good classification accuracy, and sufficiently fast to be used in hematological laboratories.

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 [1]. Automated classification of leukocytes require extracting the white blood cells from images exactly. Here the segmentation method is to enhance the leukocyte nucleus based on the combination of green color of RGB and Hue channel of HSV color space. After the leukocyte nucleus is enhanced, Otsu's multiple threshold method is applied to obtain the leukocyte nucleus from blood smear images [2]. As cytoplasm edge is colorless and unobservable, the deformable models and snake are used for cytoplasm segmentation [1]. The focus of the experiment is on the features extraction.

Rezatofighi SH et al [1] used three kinds of features, color, morphological, and texture features from the nucleus and cytoplasm areas, and two groups of texture features attained by the local binary pattern and the gray level of co-occurrence matrix are compared, but most of the features extracted are texture features, which may not have a stable performance using different dataset, besides, the time cost is longer than the manual method. D.-C [2] applied the GLCM features, and take into account of some shape features, but still the classification accuracies of lymphocyte and neutrophil are not promising. J.M. Ramirez-Cortes et al [3] present the morphological operator pecstrum, or pattern spectrum, as a feature extractor of discriminating characteristics in microscopic leukocytes images for classification, their algorithm failed in classifying the eosinophil class, and the time cost is also a little large. Ko, B.C. et al [4] extracted 12 ensemble features such as shape, color, and texture features with 71dimensions, i.e. area, perimeter and eccentricity, and geometric features, the first and second moment shape signatures, the average and standard deviation of LUV color space associated with nucleus region, and 59 LBPs features, its classification performance reach an average precision rate of 83%, which failed by using too many features without a feature selection procedure.

In this paper, we proposed a new algorithm based on shape features to classify five major types of white blood cells in peripheral blood. Firstly, after cell segmentation, morphological features are extracted from the nucleus and cytoplasm areas, and the first order moment invariant based on Euclidean distance transform [5] is extracted from the nucleus region. Then, monocytes, lymphocytes and basophils are recognized by these features, meanwhile some unknown classes which their

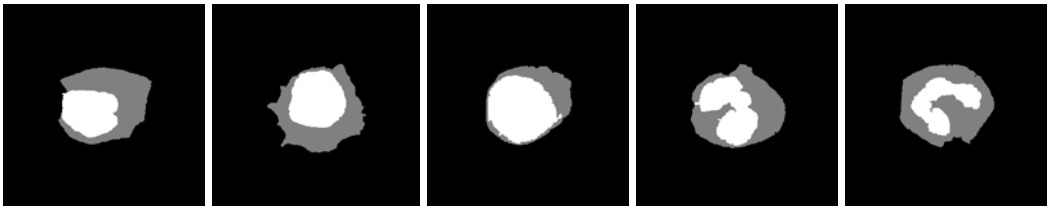
features are widely different from the major type of white blood cells are also separated. Finally, in order to classify the remaining groups of the white blood cells, the GLCM measures from the nucleus area are extracted, used as features to accurately classify via a Support Vector Machine (SVM).

Methods

Distance transformation. The Euclidean distance transform (EDT)[5], also known as distance map, computes an object point in a binary image to the nearest background points. For simplicity of representation, we assure the boundary contour of an object in an image is a continuous closed curve C. Inside contour C represents the region of the object O. The corresponding distance transform map of the object can be determined, for a point $p \in O$, calculates the shortest distance to the nearest contour point C, then transfer the distance to gray value and the distance map is achieved. In the Euclidean plane, the distance between two points (i_1, j_1) and (i_2, j_2), is given by

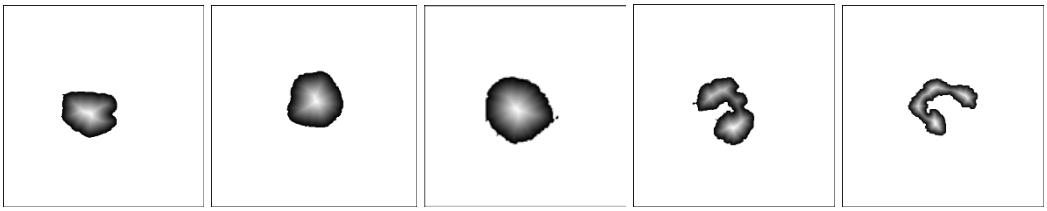
$$d_e = \sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2}. \quad (1)$$

Because the Euclidean distance transform computes on a binary image, so the image need segmented, Fig. 2 displays segmentation results of the five major white blood cells according to Fig. 1. Because of the shape of cells have an approximately round or elliptical shape, and the later experiments show the result performance of nucleus are much better than the whole cell, hence we choose the nucleus region, according to Fig. 2, the corresponding distance transform of nucleus region is given in Fig. 3.



(a) Monocyte (b) Lymphocyte (c) Basophil (d) Eosinophil (e) Neutrophil

Fig. 2. Segmentation results of the five types of white blood cells according to Fig. 1. White region is nucleus; gray one is cytoplasm; black part is the background.



(a) Monocyte (b) Lymphocyte (c) Basophil (d) Eosinophil (e) Neutrophil

Fig3Distance transform of the nucleus region of the five types of white blood cells according to Fig. 2.

Features extraction. Morphological features. In this part, morphological features used by hematologist are extracted. These features include nucleus and whole cell perimeters, nucleus and cytoplasm areas, nucleus roundness, the ratio between nucleus and cytoplasm areas, and the ratio of nucleus area to perimeter, number of separated parts of nucleus. In addition, moment invariants from the nucleus distance map are also extracted. It's used for charactering the shape of an object. The moment invariants are defined as follows.

The Geometric moment of an image is defined as[6]: $m_{p,q} = \sum_x \sum_y x^p y^q F(x,y)$, $F(x,y)$ is the function of 2D region of the image. The order of the moment is $(p+q)$.

$\mu_{p,q} = \sum_x \sum_y (x - x_c)^p (y - y_c)^q F(x,y)$ Where x_c, y_c is the center of mass: $x_c = \frac{m_{1,0}}{m_{0,1}}$ and $y_c = \frac{m_{0,1}}{m_{1,0}}$.

The normalized central moments is defined as: $\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}^{\gamma}}$, Where $\gamma = \frac{p+q+2}{2}$ The first order till 7th moment invariant is defined as follows[6]:

$$\phi_1 = (\eta_{2,0} + \eta_{0,2}), \phi_2 = (\eta_{2,0} - \eta_{0,2})^2 + 4\eta_{1,1}^2, \phi_3 = (\eta_{3,0} - 3\eta_{1,2})^2 + (3\eta_{2,1} - \eta_{0,3})^2$$

$$\phi_4 = (\eta_{3,0} + \eta_{1,2})^2 + (\eta_{2,1} + \eta_{0,3})^2,$$

$$\phi_5 = (\eta_{3,0} - 3\eta_{1,2})(\eta_{3,0} + \eta_{1,2})\left[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{2,1} + \eta_{0,3})^2\right] + (3\eta_{1,2} - \eta_{0,3})(\eta_{2,1} + \eta_{0,3})\left[3(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2\right],$$

$$\phi_6 = (\eta_{2,0} - \eta_{0,2})\left[(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,32})^2\right] + 4\eta_{1,1}(\eta_{3,0} + \eta_{1,2})(\eta_{2,1} + \eta_{3,0}),$$

$$\phi_7 = (3\eta_{2,1} - \eta_{0,3})(\eta_{3,0} + \eta_{1,2})\left[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{2,1} + \eta_{0,3})^2\right] + (3\eta_{2,1} - \eta_{3,0})(\eta_{2,1} + \eta_{0,3})\left[3(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2\right]$$

Texture features. Another kind of features commonly extracted from the object is the texture features. In this research, the gray level co-occurrence matrix is used in order to classify the eosinophil and neutrophil classes and defined as follows.

The co-occurrence matrix is constructed on a gray image with the distance d and angle θ. Assuming the number of gray level is Ng, the dimension of co-occurrence matrix is Ng×Ng [1].

Suppose the image size is M×N, and the gray level is Ng, the matrix is denoted by $W = \left[t_{ij}\right]_{N_g \times N_g}$, where t_{ij} is the $(i, j)\ th$ elements of the matrix, then a desired probability is obtained by normalizing the total number in matrix $p_{ij} = \frac{t_{ij}}{\sum_{m}^{N_g-1} \sum_{n}^{N_g-1} t_{mn}}$,Fourteen features can be extracted from the co-occurrence matrix to represent texture features. These are angular second moments, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measures of correlation and maximal correlation coefficient. To make features rotation invariants, 4 matrices and its transposed matrix are computed averaging at 4 angles.

For the remaining classes of eosinophil and neutrophil, five features are extracted from the gray level co-occurrence matrix of nucleus region to represent texture features and classifying through SVM. The five features are inertia, energy, entropy, correlation and homogeneity.

System architecture

Propose a new automatic classification of five major types of white blood cells is the main goal of our work. To achieve this goal, a block diagram is designed. Fig. 4 illustrates the block diagram of our proposed approach for classification of five types of white blood cells images.As shown in the figure, our method can be divided into two parts:

Part Ⅰ: The ratio between nucleus and cytoplasm areas, the ratio of nucleus area to perimeter and first order moment invariant are extracted to help classifying the basophil, lymphocyte and monocyte classes.

Part Ⅱ: GLCM features extracted from the region of nucleus, classifying the two remains classes through SVM.

The details of the two parts are discussed in the following sections.

In part Ⅱ, we just need to classify the two remained classes, which could be recognized easily by the nucleus features. As a result, only the nucleus features is used here. In the future, if we want to classify other unknown or abnormal cells, both the cytoplasm and nucleus features should be used. The experimental test consisted of 298 color peripheral blood cell images collected from the CellaVision reference library contain all five major types of white blood cells. First, because white blood cells consists of nucleus and cytoplasm, the performance of the first order moment invariant attained from the entire cell region and the nucleus region are evaluated, as shown in Fig. 5.

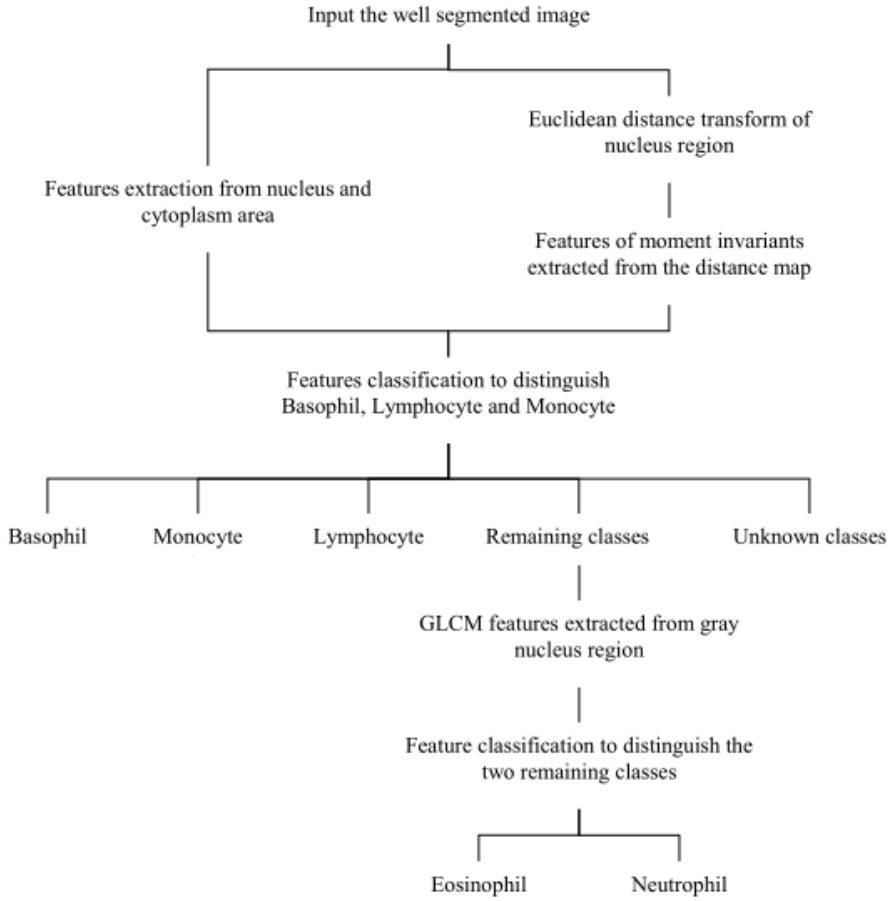


Fig. 4. A block diagram of the proposed approach for classification of five types of white blood cells.

Test results

Table 1 illustrates the confusion matrix, accuracy and overall accuracy of the classification results.

Table 1Confusion matrix, accuracy, and overall accuracy of classification results

	Recognized monocyte	*Recognized lymphocyte*	*Recognized basophil*	*Recognized eosinophil*	*Recognized neutrophil*	*Accuracy*
Monocyte	46	3	0	0	0	93.9%
Lymphocyte	3	51	0	0	0	94.4%
Basophil	0	0	64	0	0	100%
Eosinophil	4	0	0	58	3	89.2%
Neutrophil	0	0	0	0	66	100%
Overall						95.5%

According to Table 1, the result of classification of five types of white blood cells is promising, each type of cells are well classified. Table 2 summarizes the classification results of our proposed approach and other published methods.

Table 2Performance comparison results of proposed approach against other published methods

	Classification results (%)				
	monocyte	lymphocyte	basophil	eosinophil	neutrophil
J.M.Ramirez-Cortes et al. method	93.9	88.6	87.1	10.4	90.1
S.H. Rezatofighi et al. method	59.9	75.6	99.5	78.3	85.1
B.C. Ko et al. method	67.7	76.4	94.5	75.1	89.5
Proposed method	93.9	94.4	100	89.2	100

From Table 2, our approach achieves consistently high classification accuracies for all five types of white blood cells, while other methods only have a few classes well classified. Besides, the program requires only 2.1s for analyzing one white blood cell on a PC, running at 2.8GHz, with 4GB of RAM using C#. Hence, analyzing 100 white blood cells takes almost 4 min. In comparison, an expert needs about 15 min to carry out this process [1].

Conclusion

In this paper, a novel classification approach for white blood cells has been proposed and investigated. This method is based on distance transform, moment invariant, morphological features, and GLCM. We innovatively proposed three high efficiency and accuracy morphological features to separate five major cells classes into four groups. The proposed approach exhibits competitive performance compared with existed approaches, produced 95.5% overall classification accuracy, which is much better and much faster. However, the proposed method is limited to the normal leukocytes. And there is still some room for improvement in some eosinophil and monocyte. Overall, the classification accuracies produced by the proposed approach is very promising. In future research, this approach will be further improved to recognize the abnormal white blood cells and other cells which are similar to leukocytes.

References

[1]Rezatofighi SH, Soltanian-Zadeh H. Automatic recognition of five types of white blood cells in peripheral blood. Comput Med Imaging Graph(2011),doi:10.1016/j.compmedimag.2011.01.003.

[2] Huang, D.-C., Hung, K.-D., A computer assisted method for leukocyte nucleus segmentation and recognition in blood smear images, The Journal of Systems & Software (2010), doi:10.1016/j. jss.2012.04.012.

[3] J.M. Ramirez-Cortes et al. Neural Networks and SVM-Based Classification of Leukocytes Using the Morphological Pattern Spectrum. Soft Computing for Recognition Based on Biometrics, Vol. 312 (2010) 19-35.

[4] Ko, B.C. Gim, J.W. Nam, J.Y. Cell image classification based on ensemble features and random forest. Electronics Letters[J]. Volume 47, Issue 11, 26 May 2011, Pages 638-639.

[5] P. Kumar, A. Dick, Adaptive earth movers distance-based Bayesian multi-target tracking, IET Computer Vision, 7 4 (2013) 246 – 257.

[6] L. Jin, Z. Tianxu. Fast algorithm for generation of moment invariants. Pattern Recognition 37 (2004) 1745 – 1756.

[7] C.E. Honeycutt, R. Plotnick. Image analysis techniques and gray-level co-occurrence matrices (GLCM) for calculating bioturbation indices and characterizing biogenic sedimentary structures. Computers & Geosciences 34 (2008) 1461–1472.

[8] F. Nie et al. Two-dimensional minimum local cross-entropy thresholding based on co-occurrence matrix. Computers and Electrical Engineering 37 (2011) 757–767.