Research Article

A robust leukocyte recognition method based on multi-scale regional growth and mean-shift clustering

Liqun Lin and Weixing Wang

Abstract

Although there are many independent studies on the detection of white blood cell or classification of white blood cell, few papers have taken them into consideration. This study proposed a method for recognizing five types of leukocytes based on multi-scale regional growth and mean-shift clustering. The key idea of the proposed method is to extract texture features of leukocytes in a visual manner. And it is a non-parametric texture features extracting method different from traditional algorithms. Finally, SVM (Support Vector Machine) is used for classification. Some leukocyte images were used and the overall correct recognition rate reached 97.96%, indicating the feasibility and robustness of the proposed method.

Keywords

Feature extraction, leukocyte recognition, mean-shift, multi-scale regional growth, SVM

Introduction

In medical field, the analysis and identification of leukocytes are of vital importance for diagnosing diseases such as acquired immune deficiency syndrome, blood cancer, and leukemia. In particular, changes in the distribution of the five types of leukocytes (basophils (B), lymphocytes (L), neutrophils (N), monocytes (M), and eosinophils (E)) are connected with the condition of the human immune system. This analysis can be conducted using both automated and manual methods. Automated methods include flow cytometry and automated counting. These instruments can quantitatively check white blood cells (WBCs), but cannot qualitatively check them and do not benefit from image processing technologies. Applying image processing technology can provide qualitative assessment to enhance judgment. In addition, some of these tasks, such as expert manual inspection of blood cells, are tedious and prone to errors.[1,2] Thus, an automated system based on image processing technology can assist hematologists in accelerating the process. Hence, computer-aided identification methods[3–6] have been developed instead of manual methods.

Generally speaking, automatic leukocyte recognition system is mainly composed of three key steps: leukocytes detection, feature extraction, and classification. To a certain extent, correct identification of leukocytes from their background is the first step towards success. For leukocytes extraction, researchers have recently proposed a series of algorithms for accurate segmentation and classification of WBCs. For example, this work is proposed by Ghosh et al.[7] Hemocytometer provides more accurate WBCs segmentation result than manual counting, but preparation process needs expertise. Secondly, feature extraction plays a decisive role in the entire process because a set of effective features can not only compensate for the lack of segmentation but also reduce pressure on the classifier. Leukocyte features are composed mainly of geometric features,[8] histogram features,[9–11] and texture features.[12–15] The geometric features of leukocytes are effective in most cases; however, errors can be easy to make for a small number of deformed cells, such as deformed lymphocytes and eosinophils, which

College of Physics and Information Engineering, Fuzhou University, Fuzhou, China

Corresponding author:
Liqun Lin, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, Fujian, China.
Email: t07063@fzu.edu.cn

can be difficult to distinguish. In this case, they can be identified effectively by texture features. Gray level co-occurrence matrix (GLCM) and local binary pattern (LBP)[16] can be used but the appropriate parameters of two main processing methods need to be selected according to the experience; otherwise the extraction effect would be poor. Hence, determining a robust and non-parametric texture feature extraction method is necessary. Finally, as for classification algorithms, SVM,[17] artificial neural network, and decision tree are more commonly used. Zheng et al.[18] combined expectation maximization clustering algorithm and SVM for leukocyte split and classification, and this approach relies on a selected color feature vector. Zhao et al.,[19] Arslan et al.,[20] Li et al.,[21] and Zhang et al.[22] used a color space and a morphological algorithm to extract WBCs and then applied convolutional neural network (CNN) for WBC classification. CNN is applied to medical images, and the first problem is the limited training data sample.

Although the above methods of segmentation and recognition are different, the idea of the algorithms is to improve the accuracy of segmentation and classification. Based on the image color space, distance transformation, and GVF (gradient vector flow) Snake, leukocytes are extracted from cell images. Then, the author adopts a non-parametric texture algorithm based on mean-shift to extract texture feature of WBCs. Finally, SVM is used for classification.

Extraction of leukocytes

The proposed segmentation scheme is as shown in Figure 1. Based on color space, distance

transformation, and GVF Snake, leukocytes were extracted from cell images. HSI (Hue-Saturation-Intensity) color model is selected by observing different components of different color models. As shown in Figure 1, H component of HSI color model can distinguish the location of WBCs, and the effect of clustering is greatly improved.

Feature extraction

After the leukocyte is extracted, it is necessary to extract its feature. GLCM and LBP can be used but the appropriate parameters of two main processing methods need to be selected according to the experience, otherwise the extraction effect would be poor. Hence, determining a robust and non-parametric texture feature extraction method is necessary. The author adopts a non-parametric texture algorithm based on mean-shift to extract the texture feature of WBCs.

Mean-shift algorithm

Mean-shift clustering is a non-parametric algorithm that can locate the point of the maximum probability density function, which may represent a certain pattern feature. This algorithm has been applied successfully in image smoothing, image segmentation, and moving object tracking. In the d dimensional R^d space, data sampling points x_i ($i = 1 \ldots n$) are given, and the basic mean-shift vectors are defined as follows

$$M_h(x) = \frac{1}{k_n} \sum_{x_i \in S_h} (x_i - x) \quad (1)$$
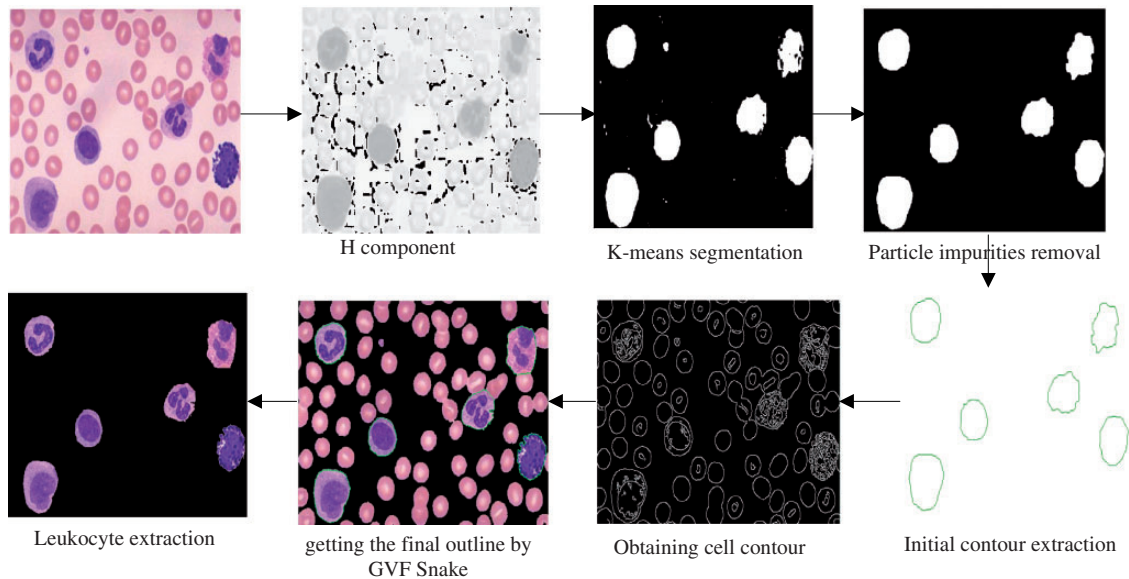
Figure 1. Flowchart of the proposed segmentation scheme.

where S_h is a high-dimensional sphere region and its radius is h, y is set to meet the following relationship

$$S_h(x) = \left\{ y : (y - x)^T (y - x) \leq h^2 \right\} \tag{2}$$

There are k_n sample points in the S_h region, $M_h(x)$ always points to the probability density gradient direction. x_i has the same contribution to $M_h(x)$ whether near or far. The different points of distance x have different weights, and hence, the kernel function $K(x)$ is introduced. The probability density function $f(x)$ can be expressed as

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3}$$

If a circular symmetric kernel is used, the profile function $k(x)$ of the kernel function satisfies the following equation

$$K(x) = c_{k,d} k\left(\|x\|^2\right) \tag{4}$$

where $c_{k,d}$ is a normalized constant to ensure $\int k(x)dx = 1$. The convergence points of $f(x)$ can be determined from their derivative zero, that is $\nabla f(x) = 0$. The gradient formula is as follows

$$\nabla f(x) = \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \right]$$
$$\times \left[\frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \right] \tag{5}$$

where $g(x) = -k'(x)$, the corresponding kernel function is $G(x) = c_{g,d} g\left(\|x\|^2\right)$, the first part of equation (5) is the probability density estimation, which takes $G(x)$ as the kernel function, and the second part (see equation (6)) is the mean-shift vector pointing to the direction of the maximum probability density gradient.

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} \tag{6}$$

Algorithm steps

1. *Selecting data space*: The color image has multiple data spaces to choose from, such as gray scale, RGB (Red, Green, Blue) color space, or HSV (Hue, Saturation, Value) color space. Selecting gray scale can be advantageous because the gray scale image can effectively reduce the adverse effects of light and WBC staining effectively.
2. *Finding the feature points*: Using mean-shift algorithm for image processing, locate the coordinates of the probability density extreme point, and obtain the gray value; however, limited data may result in some missing texture features, which require the next extension.
3. *Extending feature areas*: The texture of cell belongs to natural texture, and the change of gray level is large. Although feature points can be found, extending the feature regions accurately can still be difficult. However, the feature region can be extended using the regional growth method, but the extension is not accurate. Subsequent tests show that it can sufficiently meet the demand. The above-mentioned feature points are regional growth points, with gray scale value lower than three gray levels as the growth termination condition, and finally obtain a series of characteristic regions.

After the three steps, the processing results are shown in Figures 2 to 6. Figure 2(a) shows that the neutrophils
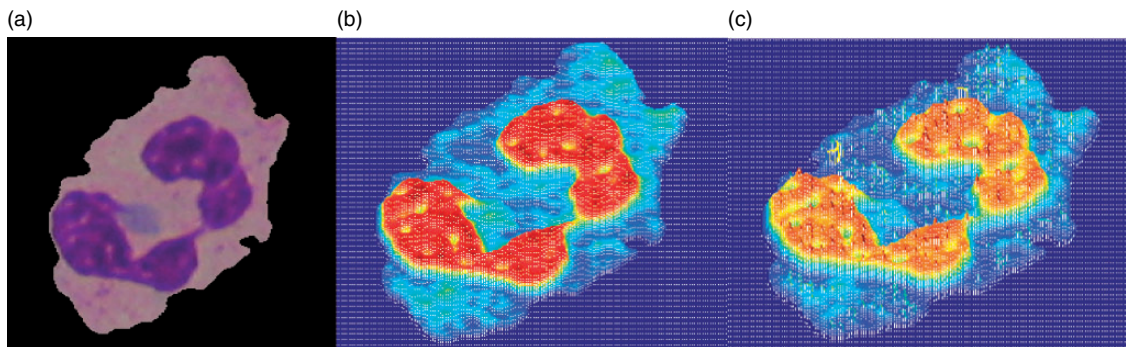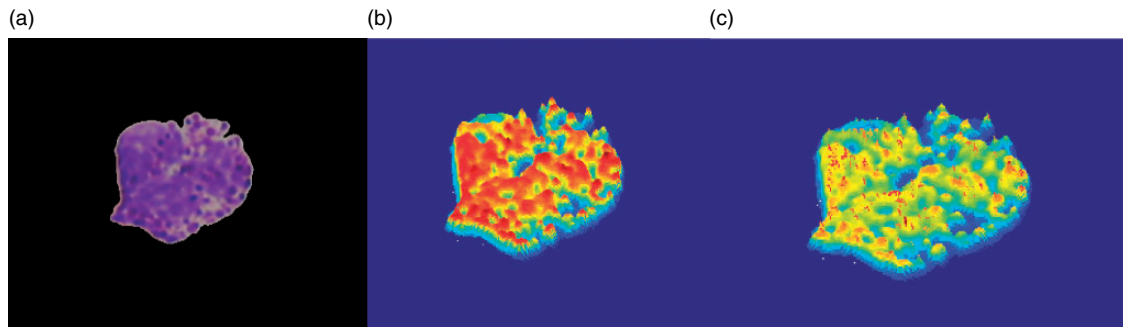

Figure 2. Processing of neutrophil. (a) Eosinophil. (b) Mesh of gray image of (a). (c) Mesh of processed image.
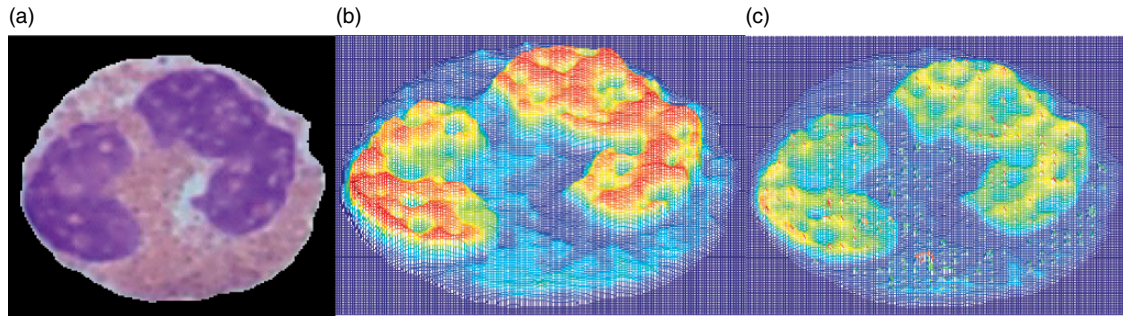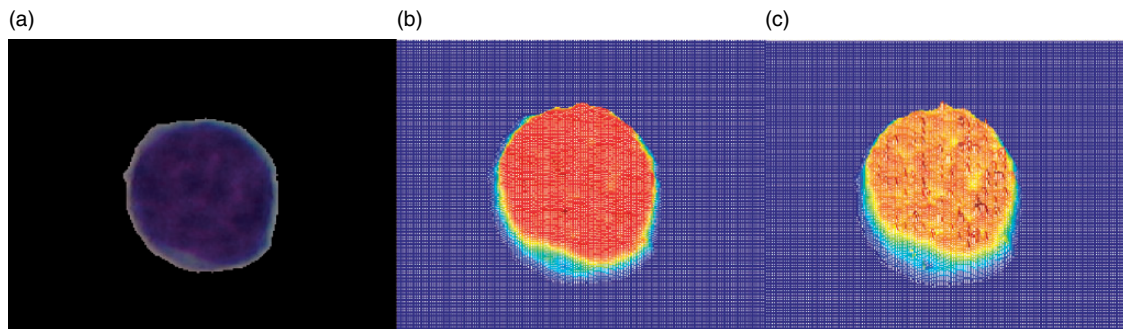
Figure 3. Processing of eosinophil. (a) Neutrophil. (b) Mesh of gray image of (a). (c) Mesh of processed image.
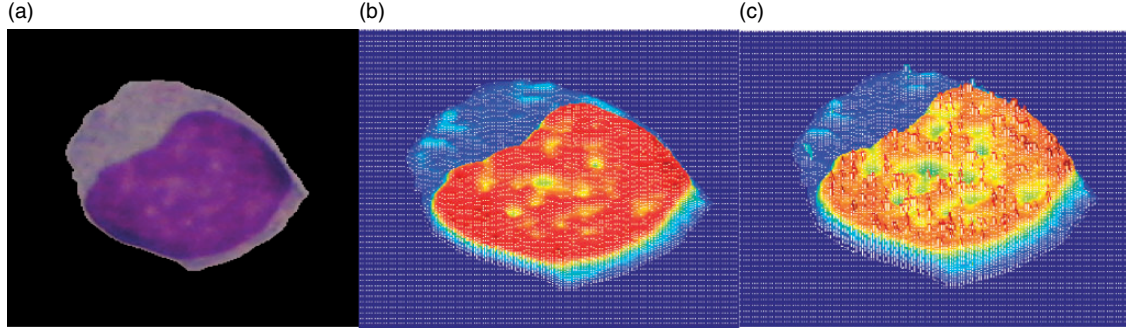

Figure 4. Processing of basophile. (a) Basophile. (b) Mesh of gray image of (a). (c) Mesh of processed image.


Figure 5. Processing of lymph. (a) Lymph. (b) Mesh of gray image of (a). (c) Mesh of processed image.

are characterized by a fine distribution of fine uniform particles in the light cytoplasm, which are also shown in Figure 2(b). Conventional edge detection and threshold segmentation may have difficulty identifying the characteristic points in the graph. After processing, Figure 2(c) shows that all the feature points are identified and the high gray value is highlighted, thereby preparing for the feature vector extraction in the next step. In the same way, after eosinophilic cells are processed by the proposed algorithm, the coarse particles that are not easy to separate from the homogeneous adhesion, can still be well labeled as shown in Figure 3(c). Figure 4 shows that the texture of the alkaline cell is not even coarse grain. The characteristic region gathers the size of the different blocks through processing. The area and uniformity of these blocks can be measured and recognized. The lymphocytes are shown in Figure 5. Compared to Figure 4(c), the regional block is relatively uniform and has a relatively small number, which could most likely ease the confusion between lymphatic and eosinophilic cells. Finally, in processing the single nucleus cell, as

(a)   (b)   (c)

Figure 6. Processing of monocyte. (a) Monocyte. (b) Mesh of gray image of (a). (c) Mesh of processed image.
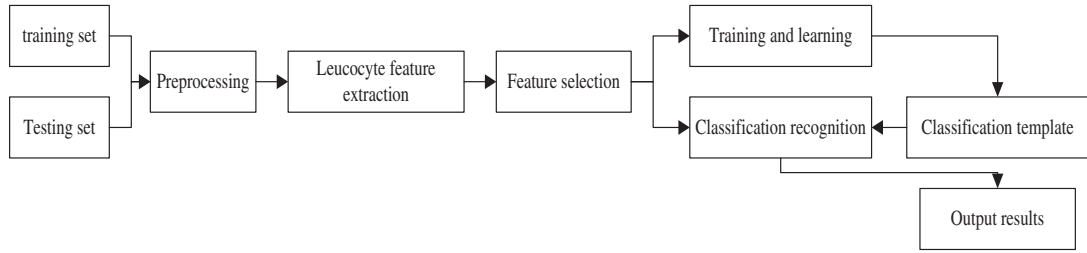
Figure 7. Flow chart of training and classmcation by SVM.

shown in Figure 6(c), its characteristic region and gray scale are small, mainly in the nucleus, cytoplasm has almost no feature points, which makes it easy to identify.

After processed by mean-shift clustering and regional growth algorithm, characteristics of five types of WBCs are highlighted. Feature regions are contracted into independent blocks, which can be distinguished by gray, area, and distribution density.

Leucocyte classification

When applying SVM for classification, a considerable number of representative WBCs are usually taken from the data to be processed as a test set. In addition, a certain number of WBCs were taken as samples for testing, and the feature was extracted from the training set, then trained with SVM classifier to get the template of classification. Finally, the classification image was classified by the classification template. The process of classification is shown in Figure 7.

Leucocyte feature extraction

1. Counting the eigenvalue and constituting feature vectors.
The feature values are as follows: average gray value of feature points ($\overline{h}_{\max}$), area of feature areas ($\overline{s}$), spacing variance of feature points ($v1$), gray point variance of

feature points ($v2$), and number of feature points (n). The variances are calculated as follows

$$v1 = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(d_i - d_{i-1})^2} \tag{7}$$

$$v2 = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(h_{\max i} - \overline{h}_{\max}\right)^2} \tag{8}$$

In formula (7), d_i represents the distance between two adjacent feature points.

2. Analyzing of feature vectors. Data in Table 1 are analyzed as follows. The mean gray values $\overline{h}_{\max}$ of the characteristic points of lymphocytes and basophils are greater than 200 because they all have dark-colored particles and high contrast. Because of its lighter color, the characteristic point average gray value $\overline{h}_{\max}$ of monocytes is less than 90, reflecting texture roughness and texture block size. The area $\overline{s}$ of the eosinophilic characteristic region is the smallest, indicating that its roughness is high and the texture block is small. The $\overline{s}$ of the monocyte is the largest, because its texture block is large and smooth. $V1$ describes

the uniformity of the texture locating distribution, and lymphocytes have a minimum, which is consistent with the human eye in terms of its compact and uniform visual experience. $V2$ can measure the color uniformity and contrast. Monocytes have the least contrast, and $V2$ has a minimum value that can be distinguished from similar lymphocytes. The number n of feature points represents the number of texture primitives, which is a good indicator of lymphoid identification. The characteristic regions are large but the number of primitives is small. Data in Table 1 show that the eigenvectors can effectively express the textures of WBCs

Table 1. Feature vector of WBCs.

Cell types	$\overline{h}_{max}$	$\overline{s}$	v1	v2	n
Neutrophil	143.30	7.75	27.61	74.65	157
Eosinophil	129.06	3.61	23.55	45.98	187
Basophil	233.40	6.47	23.81	79.05	125
Lymphocyte	223.85	7.11	18.08	54.27	65
Monocyte	80.05	12.35	25.53	22.00	211

WBCs: white blood cells.

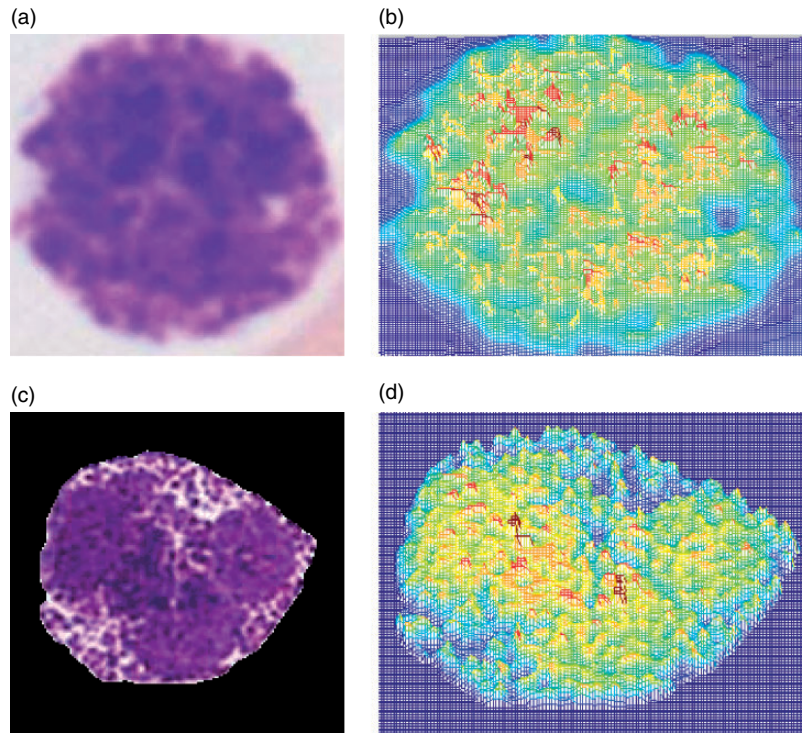and distinguish them from each other.

Robustness testing

A common recognition problem is the weakening of texture. For instance, experts identified Figure 8(a) and (c) as basophils. Compared with the standard Figure 8(c), the texture is weaker and more smooth in Figure 8(a), and thus, using the traditional GLCM and LBP, Figure 8(a) will be more easily identified as lymphocytes. The processed data in Table 2 can be identified correctly through this method. Although the $\overline{h}_{max}$ of lymphocytes and basophils are more than 200, they are very close. In addition, the characteristic point numbers n of the lymphocyte is much smaller than that of the basophil, and the $\overline{s}$ value of the lymphocyte is the largest.

Table 2. Comparison of eosinophil and lymphocyte feature vectors.

| Vector/type	Image 1	Image 2	Image 3	Image 4
$\overline{h}_{max}$	131.30	145.93	152.15	142.57
$\overline{s}$	4.32	3.56	5.37	7.09
v1	20.34	23.38	19.39	14.95
v2	59.98	46.37	70.55	51.93
n	124	134	126	70

Figure 8. Processing of different repeatability in the same type of basophil. (a) Weak texture. (b) Mesh of processed (a). (c) Standard texture and (d) Mesh of processed (c).

Note: Images 1 and 2 in Table 2 are shown in Figure 8(a) and (c), image 3 is Figure 4(a), are all basophils, whereas image 4 is Figure 5(a), which belongs to a lymphocyte.

Another situation is the small number of lymphocytes without cytoplasm, which causes them to be easily mistaken as basophils. Figure 9 shows that the n value is 57, but basophils have an n value greater than 110, and thus, they can be easily identified by the present algorithm.

Another error-prone condition is the variant lymphocyte as shown in Figure 10. Because the roundness of approximately 1 is a morphological feature of lymphocytes, deformed lymphatics are often misidentified
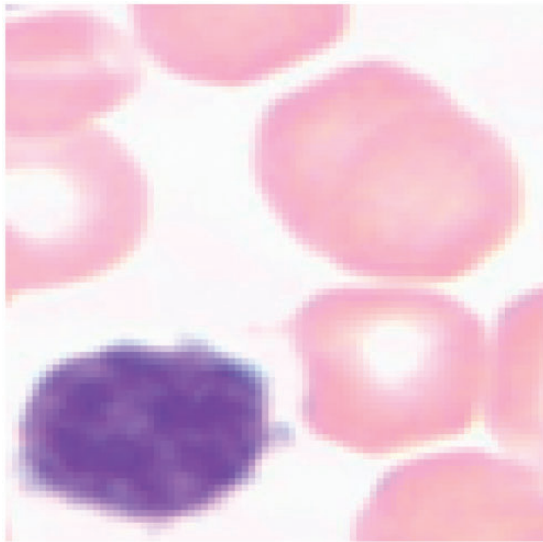
Figure 9. Special lymphocyte without cytoplasm.
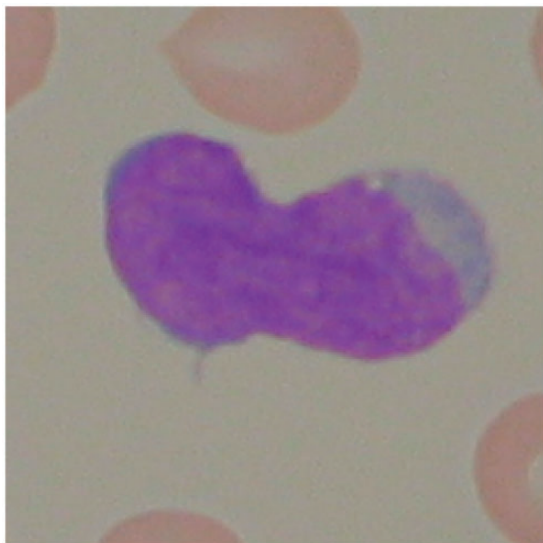
Figure 10. Deformed lymphocyte.

as non-WBCs. Using the proposed algorithm, the number n of the feature points is 38, which meets the characteristics of lymphocytes and removes the problem of being error prone.

In short, the proposed algorithm can correctly identify the conventional texture algorithm error-prone situation, so the algorithm has a high robustness.

Result and discussion

Dataset

In this paper, two sets of data were used in the experiment. The first group of data is obtained from First Affiliated Hospital of Fujian Medical University. The experimental smears were prepared through conventional methods in the hospital and stained with Wright staining. All WBCs collected from the peripheral blood and bone marrow cell images were identified and classified by laboratory specialists. In the cell image acquisition, an OLYMPUS BX51 microscope and a Nikon high-performance color digital camera were used. The blood smears were observed under microscope 100 times. The field of vision was located in the area where the WBCs were concentrated, and then transferred to the camera mode. The microscope was used to fine-tune the WBCs to the appropriate position of the image. Finally, the digital camera was used to take cell images. The second set of data is downloaded from the ALL-IDB1 (Acute Lymphoblastic Leukemia Image Database1) and ALL-IDB2 datasets.[23]

To verify the practicability of this algorithm, different types of peripheral blood and bone marrow cell images were tested and representative images selected for further analysis. The samples were pre-processed and the proposed algorithm was used to identify them. A comparison of the experimental results was also conducted.

Experimental results

Table 3 shows that after treatment with conventional morphological methods, neutrophils have significant multi-core features, resulting high recognition rate (as high as 94.6%), but the lymphocytes and basophils have similar morphologies and thus, the recognition rate was less than 83%. Morphological characteristics of monocytes were similar to the other four types of WBCs, causing the accuracy rate to drop to 90.8%.

Table 4 shows the combination of morphological features and the proposed algorithm for texture feature extraction. The recognition of the correct rate was significantly improved. Basophilic recognition rate was 100%, because of its dark grain texture was easily

Table 3. Recognition rate with geometric features.

Types	Neutrophil	Eosinophil	Basophil	Lymphocyte	Monocyte
Total amount	452	128	15	335	380
Correct numbers	428	117	12	278	345
Error numbers	24	11	3	57	35
Correct rate	94.6%	91.4%	80.0%	83.0%	90.8%

Table 4. Recognition rate with this work.

Types	Neutrophil	Eosinophil	Basophil	Lymphocyte	Monocyte
Total amount	452	128	15	335	380
Correct numbers	431	120	15	312	351
Error numbers	21	8	0	23	29
Correct rate (%)	95.4	93.8	100.0	93.1	92.3

Table 5. Objective evaluation of the classification results.

Algorithm	Lina et al.[11]	Pang and Liu[16]	Wang and Su[14]	Proposed algorithm	
Number of samples	150	150	150	150	500
Neutrophil	81.3%	96.43%	89%	96.6%	95.4%
Eosinophil	98%	100.00%	91%	100.0%	93.8%
Basophil	98%	98.64%	95%	100.0%	100%
Lymphocyte	84.3%	93.10%	97%	95.0%	93.1%
Monocyte	93.3%	95.83	96.7%	97.7%	92.3%
Overall accuracy	90.98%	96.8%	93.74%	97.96%	94.92%

extracted by the proposed algorithm and the number of samples was relatively small. Lymphatic recognition rate was also increased from 83.0 to 93.1%, because the number n of feature points serves as a good marking. The recognition rate of monocytes and eosinophils was not significantly improved, because juvenile monocytes were misidentified as lymphocytes, whereas eosinophil primary texture was very similar to that of neutrophils, resulting in higher difficulty of artificial recognition.

Table 5 shows that the correct recognition rates for basophils are similar for several algorithms, and the algorithm is optimal for the other four leukocytes. When the number of samples is increased to 500, the recognition rate decreased because of the absence of a special case, which made it difficult to identify because of small size of the sample. However, the recognition rate is still high, reflecting better robustness. The experimental data of Table 5 are based on database ALL-IDB.

Conclusion

This paper makes full use of the feature information of cell image, including its color information, gray scale information, shape and size, and distance transformation.

It also uses gradient vector flow active contour to achieve leukocyte extraction and has a good segmentation effect. And a method for extracting natural texture features based on mean-shift clustering was proposed, which is successfully used to identify and classify human peripheral blood and bone marrow leukocytes automatically. The experiment shows that the proposed algorithm has good robustness and practicability, and better recognition rate. Certainly, our proposed algorithm is not perfect. Some limitations also exist. It cannot detect all the WBCs for some complex cell images and it sometimes regards a few non-WBCs as WBCs. Adhesion segmentation for the diagnosis of complicated abnormal cells in bone marrow diseases remains a great challenge. Hence, how to find a more effective detection method based on our method is the direction of our study in the future. In addition, WBCs classification recognition can be realized by using the current popular convolution neural network, and its time cost is greatly reduced.

References

1. Saraswat M and Arya KV. Automated microscopic image analysis for leukocytes identification: a survey[J]. *Micron* 2014; 65: 20–33.
2. Dorini LB, Minetto R and Leite N. Semi-automatic white blood cell segmentation based on multiscale analysis[J]. *IEEE Trans Inf Technol Biomed* 2012; 17: 250–256.
3. Nazlibilek S, Karacor D, Ercan T, et al. Automatic segmentation, counting, size determination and classification of white blood cells[J]. *Measurement* 2014; 55: 58–65.
4. Liu Z, Liu J, Xiao XY, et al. Segmentation of white blood cells through nucleus mark watershed operations and mean shift clustering[J]. *Sensors* 2015; 15: 22561–22586.
5. Li QL, Chang L, Liu HY, et al. Skin cells segmentation algorithm based on spectral angle and distance score[J]. *Opt Laser Technol* 2015; 74: 79–86.
6. Wang Q, Chang L, Zhou M, et al. A spectral and morphologic method for white blood cell classification[J]. *Opt Laser Technol* 2016; 84: 144–148.
7. Ghosh P, Bhattacharjee D and Nasipuri M. Blood smear analyzer for white blood cell counting: a hybrid microscopic image analyzing technique[J]. *Appl Soft Comput* 2016; 46: 629–638.
8. Huanga DC, Hunga KD and Chanb YK. A computer assisted method for leukocyte nucleus segmentation and recognitionin blood smear images[J]. *J Syst Software* 2012; 85: 2104–2118.
9. Chen P, Dong SP, Sook Y, et al. Leukocyte image segmentation using simulated visual attention[J]. *Expert Syst Appl* 2012; 39: 7479–7494.
10. Habibzadeh M, Krzyżak A and Fevens T. White blood cell differential counts using convolutional neural networks for low resolution images[J]. *Artif Intell Soft Comput Lect Notes Comput Sci* 2013; 7895: 263–274.
11. Lina ■, Chris A and Mulyawan B. Focused color intersection for leukocyte detection and recognition system[J]. *Int J Inform Electron Eng* 2013; 3: 498–501.
12. Viswanathan P. Fuzzy C means detection of leukemia based on morphological contour segmentation. *Procedia Comput Sci* 2015; 58: 84–90.
13. Chastine F, Martin LT and Muhammad RW. Parameter optimization of local fuzzy patterns based on fuzzy contrast measure for white blood cell texture feature extraction[J]. *J Adv Comput Intell Intell Inform* 2012; 16: 412–419.
14. Wang WX and Su PY. Blood cell image segmentation on color information and GVF snake for leukocyte classification on SVM[J]. *Opt Precis Eng* 2012; 20: 2781–2790.
15. Syed HS, Arif IU, Saeeda N, et al. Efficient leukocyte segmentation and recognition in peripheral blood image. *Technol Health Care* 2016; 24: 335–347.
16. Pang C and Liu YJ. Improved LFP algorithm on leukocyte image texture feature extraction and recognition[J]. *Acta Photonica Sinica* 2013; 42: 1375–1380.
17. Mehdi H and Adam K. Analysis of white blood cell differential counts using dual-tree complex wavelet transform and support vector machine classifier[J]. *Comput Vis Graph Lect Notes Comput Sci* 2012; 7594: 414–422.
18. Zheng X, Wang Y and Wang G. White blood cell segmentation using expectation-maximization and automatic support vector machine learning[J]. *Data Acquis Process* 2013; 28: 614–619.
19. Zhao JW, Zhang MS, Zhou ZH, et al. Automatic detection and classification of leukocytes using convolutional neural networks. *Med Biol Eng Comput* 2016; 55: 1–15.
20. Arslan S, Ozyurek E and Demir CG. A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images[J]. *Cytometry* 2014; 85: 480–490.
21. Li Q, Cai WD, Wang XG, et al. Medical image classification with convolutional neural network. *Int Conf Contr Autom Robot Vis* 2014; 13: 844–848.
22. Zhang C, Xiao X and Li X. White blood cell segmentation by color-space-based K-means clustering[J]. *Sensors* 2014; 14: 16128–16147.
23. ALL-IDB Website, www.dti.unimi.it/fscotti/all.2015.