I.J. Information Technology and Computer Science, **2012, 1, 40-49**
Published Online February 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2012.01.06

Recognition of Marrow Cell Images Based on Fuzzy Clustering

Xitao Zheng[1], Yongwei Zhang[1], Yehua Yu[2], Jing Zhang[2], Jun Shi[2]
[1]*College of Information Technology, Shanghai Ocean University*
No.999 HuCheng Circle Road, LinGang New City, Shanghai, 201306, China
[2]*Hematology Department, Shanghai Jiao Tong University Affiliated No.6 People's Hospital*
No. 600 Yishan Road, Shanghai, 200233, China
xtzheng@shou.edu.cn, shijun7@hotmail.com

***Abstract*—In order to explore the leukocyte distribution of human being to predict the recurrent leukemia, the mouse marrow cells are investigated to get the possible indication of the recurrence. This paper uses the C-mean fuzzy clustering recognition method to identify cells from sliced mouse marrow image. In our image processing, red cells, leukocytes, megakaryocyte, and cytoplasm can not be separated by their staining color, RGB combinations are used to classify the image into 8 sectors so that the searching area can be matched with these sectors. The gray value distribution and the texture patterns are used to construct membership function. Previous work on this project involves the recognition using pixel distribution and probability lays the background of data processing and preprocessing. Constraints based on size, pixel distribution, and grayscale pattern are used for the successful counting of individual cells. Tests show that this shape, pattern and color based method can reach satisfied counting under similar illumination condition.**

***Index Terms*—C-mean Fuzzy clustering, mouse marrow, pattern recognition, cell counting**

1. Introduction

The counts of different types of white blood cells in bone marrow, the so-called differential counts, provide invaluable information to doctors in diagnosis of diseases such as AIDS, leukemia or cancers. The traditional method for an expert to achieve the differential counting is tedious, error-prone and time consuming. Analyzing the blood cell images is also important in clinics. This is especially true for the leukemia which is known as the most dangerous disease. We believe that the shape and the number of leukocytes in the bone marrow cell image are the important information for the diagnosis of leukemia. Our previous works are focused on the human marrow cell image analysis [3, 4], which can roughly get the count of each kind of cells. However, the experimental data is based on the patients' test data, and the progress of leukocyte is usually unknown before we do our visual identification. In order to keep a close tracking of the development of leukemia, we use a group of mouse to watch closely on the leukocytes developments. Methods similar to those of human beings are used to identify the different kinds of objects on the marrow sliced plates. As the staining difference and color difference, the different threshold values are used to retrieve the configuration of the cells. Similar fuzzy theory and iteration steps are followed for the whole processing. It is expected that an automatic discriminating system can be set up to save time and will let the investigation more efficient.

We expect the number of myeloblast and promyelocyte will out match the number of metamyelocyte. We also expect that the distance between these cells will be different to those later stages from earlier healthy stages. So it is important to get the cells counts of the different cells in the marrow samples. While most of these kinds of identification can be done by cell staining and then the specific color picking, our experiment has met a problem that the color can not be

Shanghai International Science and Technology Cooperation Foundation Project (11140903700).National Nature Science Foundation of China (81170507). Corresponding Author: SHI Jun, Professor, Senior Physician. E-mail:shijun7@hotmail.com

efficiently attached to those interested cells (or neucli), only one or two colors are viewable in the micro-scope image. Before we present our solution to this problem, we will cover some of the similar works done on good stained samples.

Many methods for recognition and segmentation of blood cell images have been proposed. Most of them utilized the gray level, texture, and color. In reference [1], a new, fully automated, content-based system is proposed for knee bone segmentation from magnetic resonance images (MRI). The purpose of the bone segmentation is to support the discovery and characterization of imaging biomarkers for the incidence and progression of osteoarthritis, a debilitating joint disease, which affects a large portion of the aging population. The segmentation algorithm includes a novel content-based, two-pass disjoint block discovery mechanism, which is designed to support automation, segmentation initialization, and post-processing. The block discovery is achieved by classifying the image content to bone and background blocks according to their similarity to the categories in the training data collected from typical bone structures. The classified blocks are then used to design an efficient graph-cut based segmentation algorithm. Content-based refinements and morphological operations are then applied to obtain the final segmentation. The technique does not require any user interaction and can distinguish between bone and highly similar adjacent structures, such as fat tissues with high accuracy. The performance of the proposed system is evaluated by testing it on 376 MR images from the Osteoarthritis Initiative (OAI) database. The results show an automatic bone detection rate of 0.99 and an average segmentation accuracy of 0.95 using the Dice similarity index.

In reference [2], the active contours without edges model is applied to compute the segmentation of an image into two phases. The minimization problem is non-convex even when the optimal region constants are known. The paper applies a method that can compute global minimizes by showing that solutions could be obtained from a convex relaxation. A convex relaxation approach is further proposed to solve the case in which both the segmentation and the optimal constants are unknown for two phases and multiple phases. So a relaxed convex of the popular K-means algorithm is used which can compute tight approximations of the optimal solutions.

In reference [3,4], probability and fuzzy set methods are used to retrieve cell features from image, and process automatic counting of various marrow cells that are in-sufficiently stained. Samples are based on checking records of a series of lukemia patients. The problem is that when the image quality is good, the recognization rate is good; otherwise, the algorithm will fail. To resolve this problem, higher resolution lenses are used and larger sizes of pictures are used, the expectation is that the combined image will be more efficient when the similar searching and matching algorithms are applied.

In reference [5], A SIFT algorithm in spherical coordinates for omnidirectional images is proposed. The algorithm can generate two types of local descriptors, Local Spherical Descriptors and Local Planar Descriptors. With the first ones, point matching between two omnidirectional images can be performed, and with the second ones, the same matching process can be done but between omnidirectional and planar images. Furthermore, a planar to spherical mapping is introduced and an algorithm for its estimation is given. This mapping allows to extract objects from an omnidirectional image given their SIFT descriptors in a planar image. This kind of method is useful when the current project advances to the stage that 3D data need to be processed. This is important to resolve the problem that the slicing is random and a lot of useful information can be miss-reading because of the incomplete cell parts.

In reference [6], a recognition method for the blood cell images was proposed. Since red cells, leukocytes, platelets, and cytoplasm had different color in the blood cell image, they were extracted according to their own colors. First, the color areas of red cells, leukocytes, platelets and cytoplasm were determined, respectively. Second, pixels were distributed into each color area by using the fuzzy clustering algorithm. The leukocytes, platelets, and red cells were detected accurately in all five images.

In reference [7], a method based on the color fuzzy clustering was proposed to divide the color areas and distribute each pixel into each area. The technique will segment single cell images of white blood cells in bone

marrow into two regions, i.e., nucleus and non-nucleus. The segmentation is based on the fuzzy C-means clustering and mathematical morphology. The segmentation results are compared to an expert's manually segmented images. The initial investigation of the use of the derived segmented images in the cell classification is also performed by using the Bayes classifier.

Reference [8] deals with bone marrow cell counting. An identifying and classifying algorithm is proposed using gray level and color space. An adaptive threshold segmentation method is used to analyze the HIS (Hue, Saturation, Intensity) color space of marrow cell images. H channel and S channel are developed to recognize the red cells, the nucleus and the cytoplasm of the bone marrow cells.

The morphology of the gene-perturbed cells is one of the most significant features to be examined [9]. Although morphological analysis has a significant role in medicine and biology, it does not exert its potential to the full extent. This is because most of the biomedical images contain complex structures and unclear object boundaries and thus modern image processing and computer vision algorithms still cannot analyze them successfully. E.g., the success of the yeast morphological study depends on highly-controlled clear images of isolated cells. In general, the automatic morphological analysis of biomedical images is a challenging problem, and there is a strong demand to solve this problem.

Reference [10] presents the similar work done for teeth image recognition, a fuzzy neural network algorithm is used to process fish counting, which has a simple pattern set to simulate various fish position. Reference [11] proposes a fuzzy membership function to precede the ambiguous part matching of human teeth. Here will exploit the features of different components of marrow samples using a small program, get the statistical features of the interested objects, and then search these objects in the original images to enable the automatic object recognition. Due to the size of the paper, only the first part, the featuring and identification of leukocyte will be introduced.

So most previous methods followed the traditional manual maneuver for blood cells, i.e., detecting a cell, extracting its features, classifying the cell, and then updating the count. Even though several attempts have been made to solve the marrow cell counting, they can only be applied to specific blood samples. The counting problem in bone marrow is much more difficult due to the high density of cells. Moreover, there are many types of bone marrow white blood cells that may not be found in the blood. Our previous works were all applied to the counting of leukocyte. In many cases, only nucleus information is adequate to classify a cell.

2. Methodology

The fuzzy c-means (FCM) is a method to cluster objects which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. In our practice, in order to develop a theory that can combine the mathematical morphology and membership function, we need to develop a new training algorithm for membership function in order to pick out objects from different cell classes. Here we will combine the common techniques used in cell segmentation like thresholding, cell modeling, filtering and mathematical morphology, clustering and fuzzy sets.

The fuzzy C-means algorithm (FCMA) is often used to segment cell images. We will briefly introduce it and the related membership function construction. The mathematical morphology which covers the pattern retrieving and representation will not be elaborated here. We have used this method to process the human marrow cells and the result is acceptable.

extensive coverage of this topic.

A) Fuzzy C-Means Algorithm (FCMA)

We start from the normal fuzzy clustering method. Let us assume that the expression of the set of factors that affects the evaluation is U, the set of n comments is V. Here the comments represent the level of the acceptance to membership function.

We use matrix R to represent the relation between the factors domain and the comments domain:

$$R=[r_{ij}]\ (1\leqslant i\leqslant m,\ \ 1\leqslant j\leqslant n) \quad (1)$$

In the full evaluation of one object, all the factors should be considered. The weight of every factor distributed by the evaluator can be represented by

$$A=(a_1, a_2,\cdots, a_m) \quad (2)$$

Where $\sum a_i=1$, $0<a_i<1$, i = 1,2,⋯m.

The composition of A and R can be thought as the final evaluation of the objects, or the fuzzy comprehensive evaluation. The mathematical model is presented as the following expressions.

$$B = A \bullet R = (b_1, b_2, ..., b_n) \quad (3)$$

In the expression (3), B or b_j represents the membership grade of the evaluated object.

As in c-mean fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. This indicates the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

Here we use the max-min similarity relation to resolve the equivalence relations. The max-min similarity-relation matrix was made up of correlation coefficient R or r_{ij} as in formula (1), which was obtained by the max-min formula as follows:

$$r_{ij} = \frac{\sum_{k=1}^{m} x_{ik} \wedge x_{jk}}{\sum_{k=1}^{m} x_{ik} \vee x_{jk}} \quad (4)$$

When the data set can not be represented by formula (1) to (4), but can be divided into several parts to finish the same, we call this C-means clustering. This way the data can be evaluated cluster by cluster and the clusters are also co-related and co-evaluated.

Consider a set of data X = {x_1, x_2,..., x_n}, where x_k is a vector. We would like to partition the data into c clusters. Assuming that we have a fuzzy pseudo partition P = {A_1, A_2,..., A_c}, where as introduced in part A, A_i contains membership grades of all x_k to cluster i. The centers of the c clusters can be calculated by

$$v_i = \frac{\sum_{k=1}^{n} [A_i(x_k)]^m x_k}{\sum_{k=1}^{n} [A_i(x_k)]^m}, \quad i \in (1, C) \quad (5)$$

Where m > 1 is a real number that controls the effect of membership grade. In the FCMA, the membership grade of the vector x_k to cluster i is defined as follows: if $||x_k - v_i||^2 > 0$, then for all i ∈ {1, 2,...,c}, we have

$$A_i(x_k) = \left[\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|^2}{\|x_k - v_j\|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (6)$$

$A_i(x_k)$ follows the same rule as formula (3)

The performance index of a fuzzy pseudo partition P is defined by

$$j_m(P) = \sum_{k=1}^{n} \sum_{i=1}^{c} [A_i(x_k)]^m \|x_k - v_i\|^2 \quad (7)$$

The clustering goal is to find a fuzzy pseudo partition P that minimizes the performance index $J_m(P)$. When the sub-partitioning is easier to be expressed, the C-mean is simpler than multilayer clustering.

We need to point out that the treatment from formula (5-7) is for calculation purposes only and the normal fuzzy clustering maximizing method will also work here. Normal fuzzy clustering maximizing method will also work here.

B) Clustering of objects

In our previous work to process human marrow cells, objects are classified into 6 categories and these clusters can be used to identify red cells from leukocytes, which are dependent on the pre-selection of C1 to C3. C5, the pixel distribution pattern, can be quite similar or different depending on the mapping area selection, i.e., the border can significantly change the pre-trained distribution. So it is hard to identify different leukocytes only using C1 to C6 in Figure 1. C7 and C8 [4] are introduced over the original sets, which are the ratio of neucli to cytoplasm and the relative sizes. The meta-myelocyte and later phases will be considerably smaller in size and have lower neucli-cytoplasm ratio.

The images are traversed with the windows (circles) sized by the value of C6 (an array varies by objects), where C5 (also an array) are calculated. C1 to C4 can also be trained for a specific batch. A combination of C1 to C6 servers as V in formula (3-4), and the actual figures, which are from the traversing actions over the whole image, will be X. The matrix is huge and the elaboration of the application detail may not be allowed by the size of this paper. This first round traversing will mark off all the red cells, megacaryocytes, and the leukocytes. All the leukocyte will be copied to a new

image and processed with C7 and C8 clusters to get the metamyelocytes from leukocytes. I,e, the C-mean will be performed again over the two sets to get the minimum J_m (P).

So for each loop, the test window is moving over the image, J_m(P) will be calculated for each step over the designated object sets and the quasi-minimum value will be compared to determine if it is the target.

For the mouse cells, similar algorithms are applied to the searching and clustering practice. When proper threshold values (RGB values) are selected and correct edge detection methods are applied, the algorithm works fine and 90% or more recognization rate can be yield. But problem is the image quality is not stable and the color threshold needs to be selected by experienced operators.

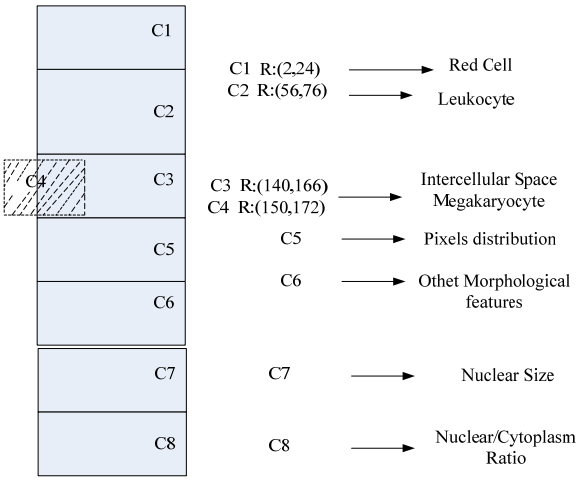

Fig 1 Object clusters and their relations

3. Experiment

The experiment is done over a 15 days period, totally 15 mice are used for the process. All the mice are injected lukemia inducing fluid and are killed daily to get the slicing of its femur. Both the middle and end of the bones are selected as the watching zones. The samples are stained and fixed and sliced into thin plates and placed under microscope for observation. The amplification rate is 400 folds. So considerable amount of data are available for processing.

As we know, the majority of the jobs for FCMA are for well-stained samples, in order to use it on our samples, we need to carefully group the clustering sections so that the different groups can be differentiated at the minimum J_m (P). Our grouping details can be seen from Fig 3, when the sections boundaries are carefully selected using trial-and-error method and the values are also batch-related.

Our experiment is based on multiple images of 717*516 pixels, part of which can be shown in Fig 2. Fig 3 shows the interested targets for our identification process, parts of their features are shown in fig 4. There is only one megakaryocyte in this image, so only one sample is listed. The leukocyte has 5 samples and the red cell has 3. This is because the leukocyte is our key interest and will be further processed in the program.

Smaller images are selected to easily show effectiveness of the algorithm, although working on this kind of smaller images will make the result more vulnerable because the base cell number is too small. Dozens of pictures from the same staining batch have been processed and the results are compared with human marking, the average recognition rate stabilizes at 80%-90%.

The experiment has two parts, one part is the learning program and another part is the recognition program. In the learning program the user will be able to select an object type, and then pick a few objects of this type from the screen, the computer will learn this object and decompose the objects into the x_k array, and the way to process different type of cells will be different. To process megakaryocyte, morphological features will be extracted and stored. To process other cells or spaces, gray value distribution and pixel distribution will be extracted and stored. These data will be classified into four clusters and stored for recognition matching.

The recognition process will be performed in the following sequence: the search of megakaryocyte, the search of red cell, the search of leukocyte, and probably, the search of promyelocyte. The first two steps will include a mark off program so all the identified part will be marked off, so the last step will be leukocyte and intercellular space only, which can be easier to apply pattern features.

The identification proved to 95% or more with megakaryocyte, 80% or more with leukocytes and red cells. The identification rate of promyelocyte varies on the method used. The statistical features go at lower rate and the morphological feature goes at a higher rate. This

is because the promyelocyte has a bigger size. Because of the poor staining, blurred image and the similar shape, the identification algorithm for leukocyte still need to be improved.

The experiment uses a window of 100*100 square pixels for metamyelocyte cell detection, 50*50 square pixels for leukocyte cell detection and 40*40 square pixels for red cell detection. This size can be different depend on the resolution of the image and the enlarge rate of the image, and can be preset during the learning process. Additional pictures can be fed without further treatment if the first image is correctly processed.

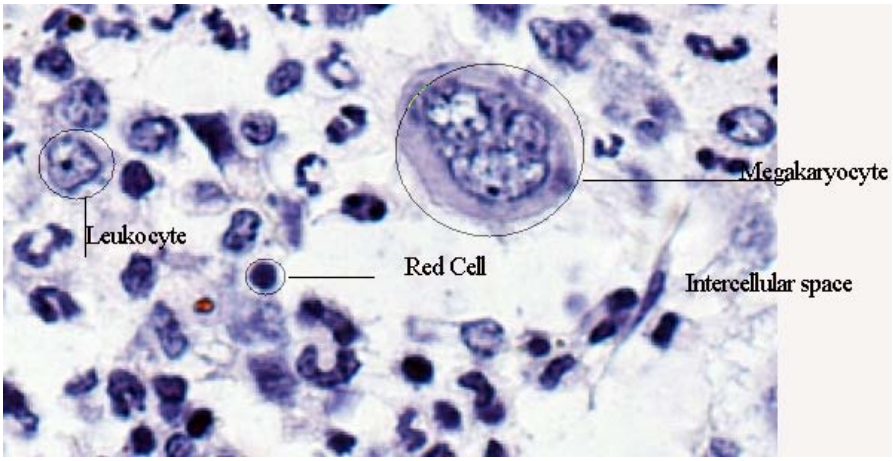

Fig 2 Part of experimental image and its illustrations

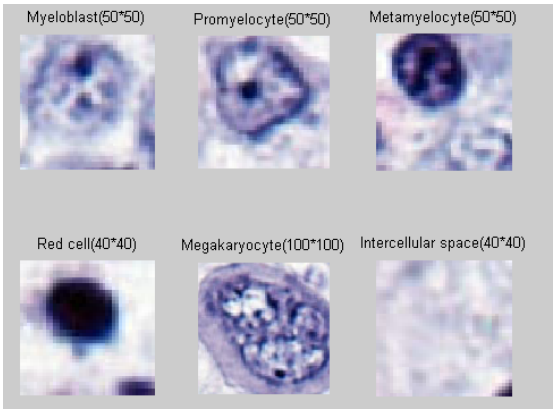

Fig 3 The elements that need to be extracted

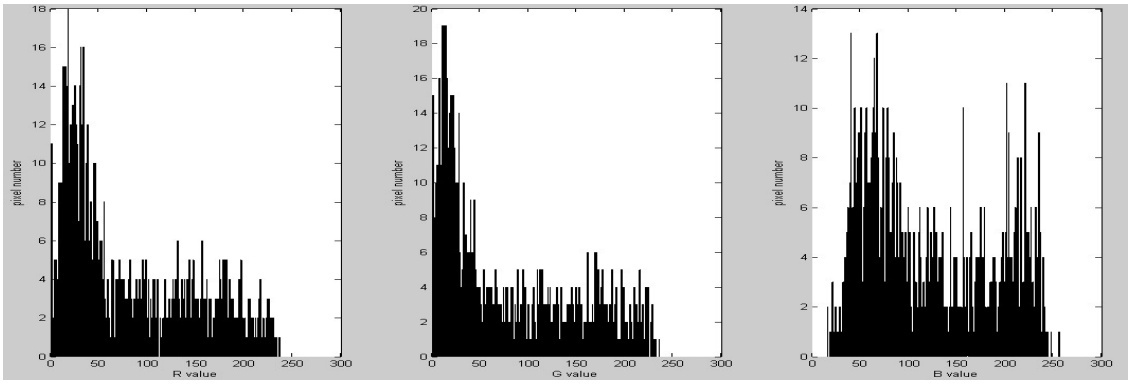

Fig 4.1 pixel-gray value chart of intercellular space RGB

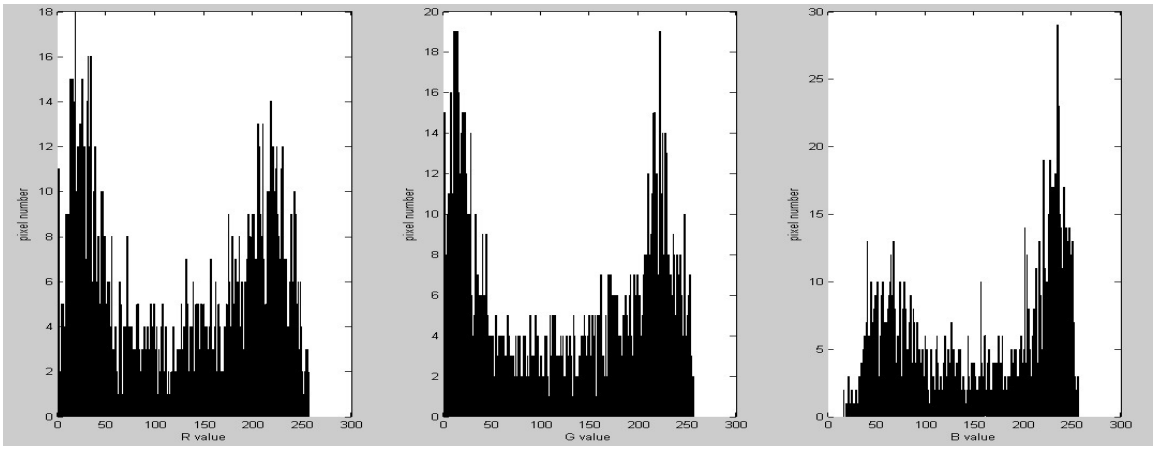

Fig 4.2 pixel-gray value chart of leukocyte space RGB

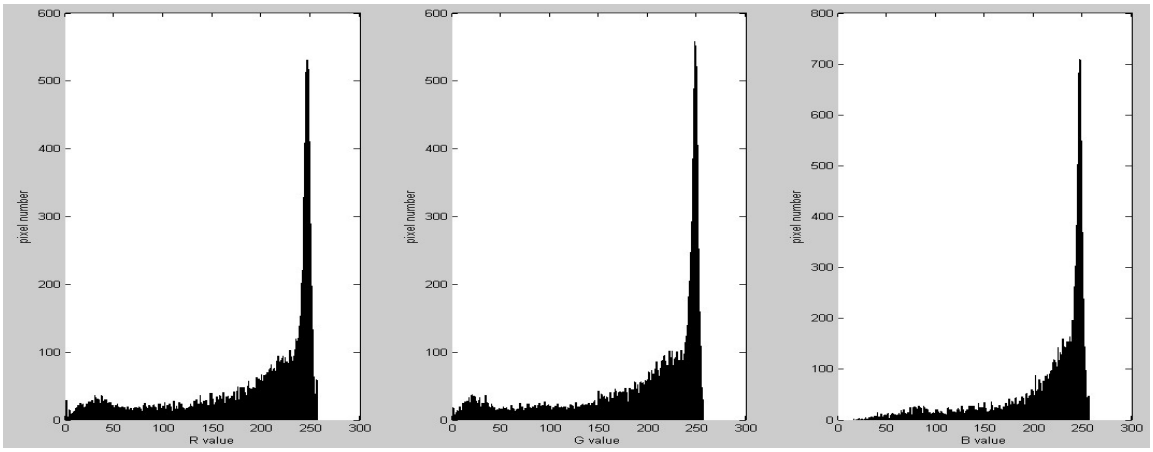

Fig 4.3 pixel-gray value chart of megakaryocyte space RGB

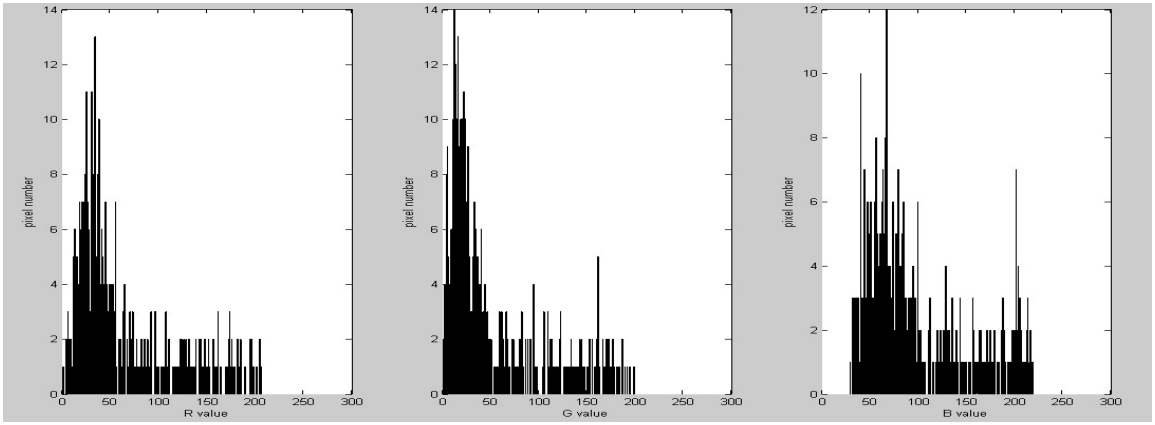

Fig 4.4 pixel-gray value chart of red cell space RGB

From Fig 4.1—4.4 we can find different grayscale distributions for different cell targets, these distributions are stable compared with the kinds of cells. The threshold value to segment the picture and turn the picture into grayscale can be seen in Table 1, which is also a range depended on the batch of image.

Table 1 pixel distributions over RGB values

	R distribution	G distribution	B distribution	Cell size (pixel)
Leukocyte 1	122.6	123.1	157.9	37*37
Leukocyte 2	131.8	132.7	164.4	42*35
Leukocyte 3	107.4	107.6	146.1	40*40
Red cell 1	62.4	52.4	98.6	19*21
Red cell 2	61.8	52.0	98.1	20*21
Red cell 3	62.7	53.6	101.1	21*19
Megakaryocyte1	189..5	189.9	209.1	107*121
Megakaryocyte2	189.7	190.4	209.0	86*112
Intercellular space1	67.1	63.1	109.3	26*30
Intercellular space2	61.5	54.5	100.2	24*27

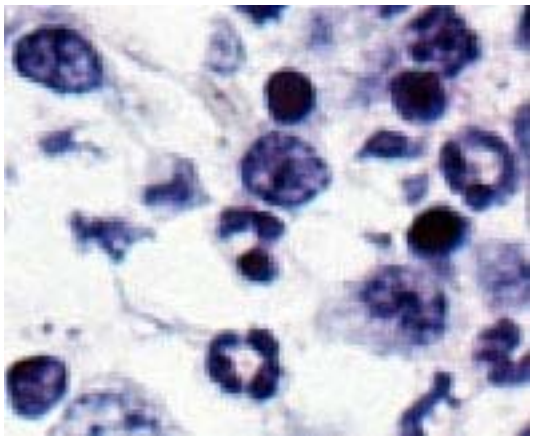

Fig 5.1 Sample (original)

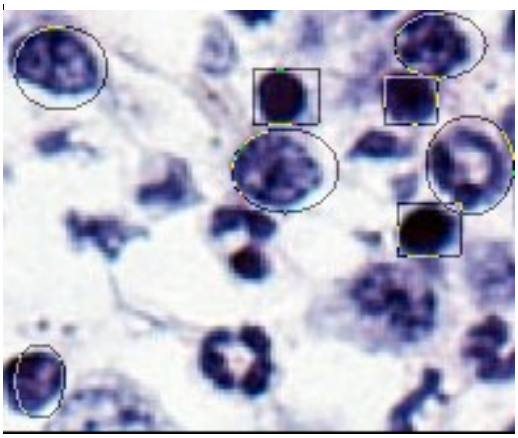

Fig 5.2Sample (expert)

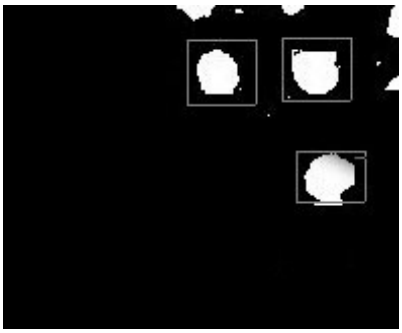

Fig 5.3 Red cell (rectangle), upper

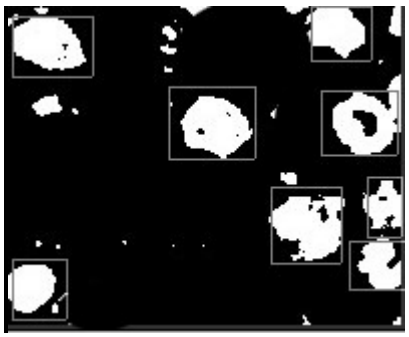

Fig 5.4 Leukocyte (rectangle), lower

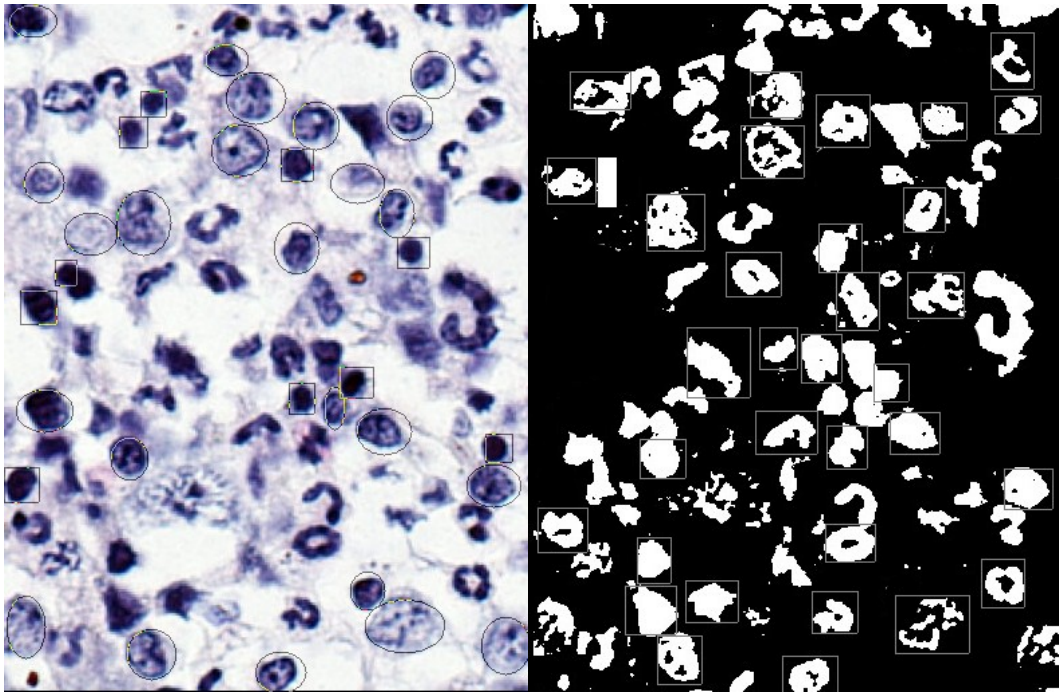

Fig 5.5 Sample (expert) upper

Fig 5.6 Leukocyte (program) lower

Fig 5.1 is the original image and Fig 2 the expert picks. Fig 3,4 are the program picks and Fig 5.5,6 are the comparison with another example of the comparison. These counting results are listed in Table 2 and it can be seen that in most case the matching is good.

For better recognization rate, the program can be run on different combinations of threshold so the final count can be more reasonable, this will be in cooperated with cell position records and multiple matrix processing and is not covered in this paper.

Table2 Leukocyte and Red cell recognition comparison

		Leukocyte	Red cell
Expert count	Sample1	5	3
	Sample 2	20	9
Program Count	Sample 1	8	3
	Sample 2	28	14

As an example, data from table 1 and 2 are the key features to be used for C1 to C6 clusters, usually C5 and C6. The images are traversed with the windows sized by the value of C6 (an array varies by objects), where C5 (also an array) are calculated. C1 to C4 can also be trained for a specific batch. A combination of C1 to C6 servers as V in formula (5-7), and the actual figures, which are from the traversing actions over the whole image, will be X. The matrix is huge and the elaboration of the application detail may not be allowed by the size of this paper.

For sample1, red cell and leukocyte can be separated by these two tables. When the test window is moving over the image, $J_m(P)$ will be calculated for each step and the quasi-minimum value will be compared to determine if it is the target.

We processed hundreds of pictures for different staining batches, the result is listed in the table as we stated above. To illustrate the details, more samples are laid out here as Fig 5-7, where small pictures and big pictures are presented. If we are doing the multiple cell recognization parallelly, we may not be able to differentiate some overlapped cell. But if we are doing the calculation sequentially, we can screen the bigger one first, then mark the position off the image, this way the recognition rate can be increased by 2-5%, and different stages of leukocyte can also be achieved by adding more evaluating factors.

From Table 3 we can see that from the different samples the failed rate is around 20%. This is because the slicing and imaging problem which may get a small part of a cell or the staining is poor so the imaging for some area will be blurred.

We need to point out that the recognition rate of cells by image is dependant on the imaging condition, like the illumination and staining condition, training is necessary for each batch. The machine sometimes provides better result because of the image processing.

From Table 1 and 2 we can see that the computer picks are mostly conformed with expert picks. We processed hundreds of pictures later and found that the result is stable for one staining and picturing batch, change of staining and photoing conditions will increase the burden of training and experience is required when picking the training area.

4. Conclusion

The C-mean fuzzy set plus geometrical extraction and sequential searching method can successfully identify object from the mouse marrow cell images. Those images can be in RGB color models. The methods divide the 6 types of target into 8 clusters with specific features and finally combine the cluster features to achieve successful object retrieval. The cluster classification is unique to the mouse marrow cell image and will change depending on the staining and imaging condition. The cluster C7 and C8 can be used to after the pre-selection of leukocytes to get metamyelocyte from leukocyte. The object positioning is not covered in this paper. Better recognition rate can be achieved using combination of high resolution pictures, repositional system from high to low resolution picture is required for the leukemia evaluation system.

References

[1] Sufyan Y. Ababneh, Jeff W. Prescott, Metin N. Gurcan. Automatic graph-cut based segmentation of bones from knee magnetic resonance images for osteoarthritis research. Medical Image Analysis 15 (2011) 438–448

[2] Ethan S. Brown , Tony F. Chan , Xavier Bresson. Completely Convex Formulation of the Chan-Vese Image Segmentation Model. INTERNATIONAL JOURNAL OF COMPUTER VISION.2011.

[3] Xitao Zheng, Jun Shi, Yehua Yu, Yongwei Zhang. A New Method for Automatic Counting of Marrow Cells. Proceeding of the 4th International Conference on Biomedical Engineering and Informatics, 2011, page 44-48.

[4] Xitao Zheng, Jun Shi, Yehua Yu, Yongwei Zhang. Analysis of leukemia Development Based on Marrow Cell Images. Proceeding of the 4th International Congress on Image and Signal Processing, 2011, page 95-99.

[5] Javier Cruz-Mota, Iva Bogdanova , Benoît Paquier, Michel Bierlaire, Jean-Philippe Thiran. Scale Invariant Feature Transform on the Sphere: Theory and Applications. INTERNATIONAL JOURNAL OF COMPUTER VISION.

[6] En-yong Wang,Zhengpin Gou, Ai-min Miao, Shu-qing Peng, Zhen-yang Niu, and Xin-lin Shi. Recognition of Blood Cell Images Based on Color Fuzzy Clustering. Fuzzy Information and Engineering Volume 2. Advances in Soft Computing, 2009, Volume 62/2009, 69-75

[7] Nipon Theera-Umpon. Patch-Based White Blood Cell Nucleus Segmentation Using Fuzzy Clustering. *Transactions on Electrical Eng, Electronics, and Communications*, ECTI-EEC 3, 15-19 (2005).

[8] AI Da-Ping, YIN, Xiao-Hong, LIU Bo-Qiang, LIU Zhong-Guo, YUAN Qing-Wei, LI Xiao-Mei. The Algorithm of Marrow Cell Identification and Classification. Chinese Journal of Biomedical Engineering, 2009, 28(4).

[9] Hiroshi Hatsuda. Automatic Cell Identification Using a Multiple Marked Point Process. *The 2010 International Conference on Bioinformatics and Computational Biology.*

[10] Xitao Zheng, Yongwei Zhang . A Fish Population Counting Method Using Fuzzy Artificial Neural Network. *The 2010 International Conference on Progress in Informatics and Computing conference*, page 225-228.

[11] Junlan Shang, Xitao Zheng, Yongwei Zhang. A Teeth Identification Method Based on Fuzzy Recognition. *The 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2010, page 271-275.

