Blood Cells classification by Image color and intensity features clustering

R. A. Melnyk[1], A.O. Dubytskyi[2]

[1]Doctor of Technical Science, Professor, Institute of Computer Science and Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine; E-mail: ramelnyk246@gmail.com

[2]The 2nd year master, Institute of Computer Science and Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine; E-mail: andrii.dubytskyi@gmail.com

Abstract – A new approach for cells detection and classification on blood smear images is considered. Benefit of 4-connected over 8-connected component labeling for cell detection is shown. Color and intensity histogram clustering are proposed to extract common features for cells classification. A new approach for k-means initial centroids detection proposed. The algorithms effectiveness was tested and estimated for some blood smear images. The algorithm examples, figures and result table to illustrate the approach are presented.

Key words – computer vision, visual object detection, visual object classification, binarization, connected component labeling, intensity feature, color feature, cluster analysis.

I. Introduction

Blood smear analysis is an important task in desease diagnostic. Complex and expensive hardware solutions are often used nowadays, and they require such an expensive reagents. That is why, fast automatic analysis of blood smears with minimal costs and good precision are needed by a labs all over the world. Computer vision algorithms applied for blood smear images can solve this task. Blood smear analysis from computer vision standpoint can be divided into two stages: cells detection and cells classification. Desease classification is related to medical knowledges and is not tangent with computer vision issues. Image processing can be applied before each of these steps.

There are two popular modern object detection approaches: sliding-windows and segmentation. [1], [2] The main drawbacks of sliding-windows are complexity and detection difficulties in case of large intra-class variation. Because of a big number of window positions and size combinations, and feature extraction on each window, even if it does not contain the object, it might be very complicated to perform object detection from computational standpoint. Segmentation approach on the other hand by itself represents similar objects as one global region and it could be very difficult to extract and count separate objects. In this article, we propose to use connected-component labeling [3] over the binarized image to detect separate cells. We propose to use Otsu thresholding algorithm [4] for binarization purposes.

Color and intensity features are commonly used for visual object classification. We propose to apply k-means clustering [5] for features to train classifier, then calculate distance between randomly selected cell feature and centroids ontained during training. Intensity feature is commonly used for classification purposes in differen computer vision approach, but blood cells has many intra-class variations and it might be complicated to classify it only by intensity feature. We propose to use intensity and color features simultaneously. Color space feature by itself represented by millions of color shades, requires a huge amount of memory and can lead to a great delays while calculating distance between features. Color space quantization [6] can take place to solve this issue.

An example of original blood smear image for experiment purposes shown in Fig. 1.

II. Image Processing

Usually computer image is a matrix of pixels each representing millions of different color shades. In this case it should be very difficult to detect different objects from the image. Some algorithm that should separate object related pixels from background pixels is required. We propose to use Otsu thresholding algorithm for this reason. Otsu's binarization defines a global threshold for color intensities histogram. After binarization, pixels with the higher intensity then threshold can be considered as background and pixels with the lower intensity can be considered as objects. We have performed a series of attempts to apply local regions detection and perform thresholding in those regions. This experiment shows that local thresholding does not lead to satisfied binarization quality improvement, but causes performance loose. However, manual threshold control can improve binarization quality. Increasing threshold causes quality loss, decreasing threshold on the other hand causes to higher quality and better detection results. Result of image binarization shown in Fig. 2.

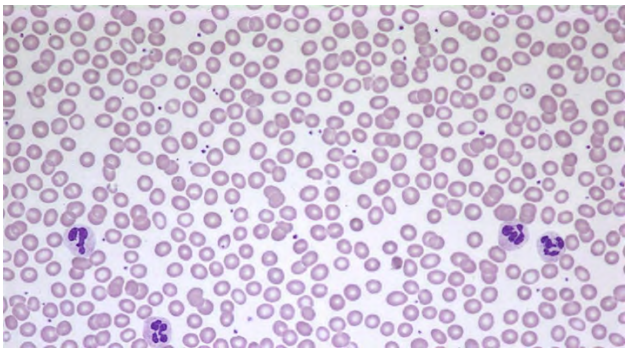

Fig. 1. Granulocytes circulating in the blood of a patient with a normal peripheral smear [7]

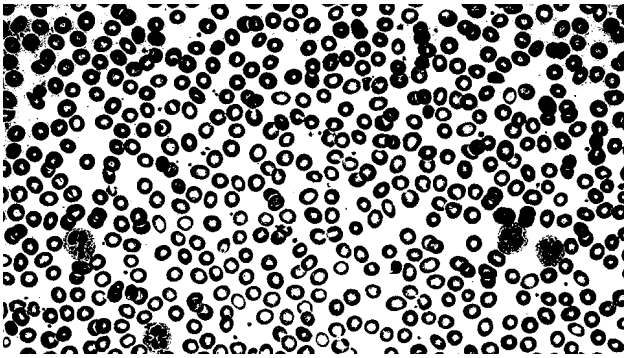

Fig. 2. Binarized blood smear image

Different blood smear images may have different resolution, which can lead to the higher distances between histograms during classification and as a result incorrect classification results may be obtained. Thus, we need to ensure scale invariance by zooming cell image to some pattern size either while training classifier or classify cells.

III. Cell Detection

We propose to use connected component labeling over the binarized image for cell detection purpose. This approach will help to obtain regions that corresponds to separate objects with minimal time costs. Thus, we will not need to extract features from regions that are not objects as in sliding-window approach.

There are two options for connected components labeling: 4-connectivity and 8-connectivity detection. We observed that 4-connectivity detection gives results that are more precise for current problem field. Result of component detection with different connectivity and binarization threshold correction shown in Table 1.

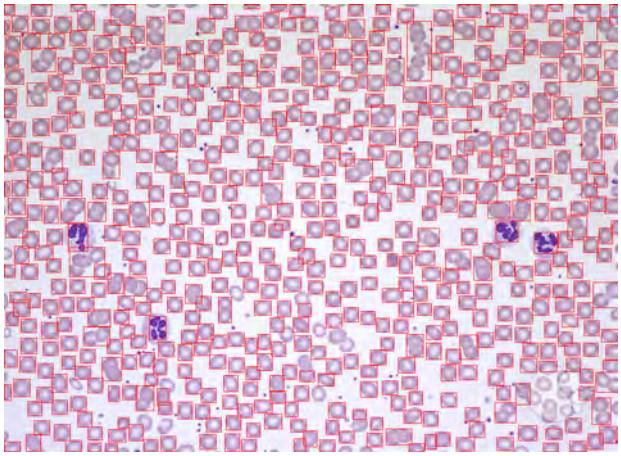

Fig. 3. Detected cells

However, there are two issues.

First problem is that the cells usually has a center with higher intensity then threshold, and those pixels will not be included into the connected component. For further analysis, we need to retrieve original pixels that corresponds to connected component's elements. That is why we need to include those cell's centers into components to perform proper classification. Another problem is that there may be a lot of trash element on the blood smear image, such as parts of dead blood cells. To solve the problem of trash detecting we need to consider a maximal possible size of trash components and remove all components that have less or equal sizes. For this reason, we perform average component's size calculation during component detection, then multiply it with a coefficient and consider resulting value as the maximal possible size of trash elements. Moreover, we need to find such coefficient that will not cause to removal of real cells. According to the experiments the optimal size of trash is the 65% of average cell size, thus the coefficient should be 0.65. Cell detection results due to the optimal algorithm parameters are shown in Fig. 3.

The algorithm that fills the spaces in components looks as follows:

```
for each row in matrix do
 set currentComponent to 0
 set currentComponentColumn to -1
 for each column in matrix do
  if matrixElement > 0 then
   if currentComponent = 0 then
    set currentComponen to matrixElement
    set currentComponentColumn to column
   else if currentComponent = matrixElement then
    for i = currentComponentColumn to column+1 do
     set matrix[row][column] to currentComponent
    end for
   else
    set currentComponent to 0
    set currentComponentColumn to -1
   end if
  end if
 end for
end for
```

IV. Feature Extraction

Image feature extraction are usually applied for classification purpose. We propose to use intensity and color features simultaneously. Both this features can be visualized as histograms.

Color feature represents the distribution of colors in an image. Nowadays colors howevers are usualy represented by millions of different color shades. That makes color feature storage and processing very complicated task. Color space quantization should be applied to minimize color space. All amount of colors can be distributed into 64 quants by partition each color chanel into 4 equal parts and forming 64 color cubes.

Intensity feature represents the distribution of color intensities over the image. Majority of approaches uses the quantization of this intensity values into 16 values, but we propose to take into consideration all of 255 intensity values. It is not so much, so it will not affect performance very much and it can give us more precise feature characteristics. To ensure scale invariance statistical normalization of histogram values should take place.

V. Feature Clustering

For visual object classification purpose we have to extract features for separate cell types. However, most of the cells of the same type on the blood image can vary by its histograms in some range. Moreover some cells, like leukocytes, have different subtypes, which are not very different. Common feature extraction for corresponding types is required to provide proper classification. We propose to use k-means clustering for common features extraction. Obtained cluster centroids after cluster analysis procedure can be considered as common features for different cell types. Some common intensity features for different blood cels are depicted in Fig. 4, common color features – in Fig. 5.

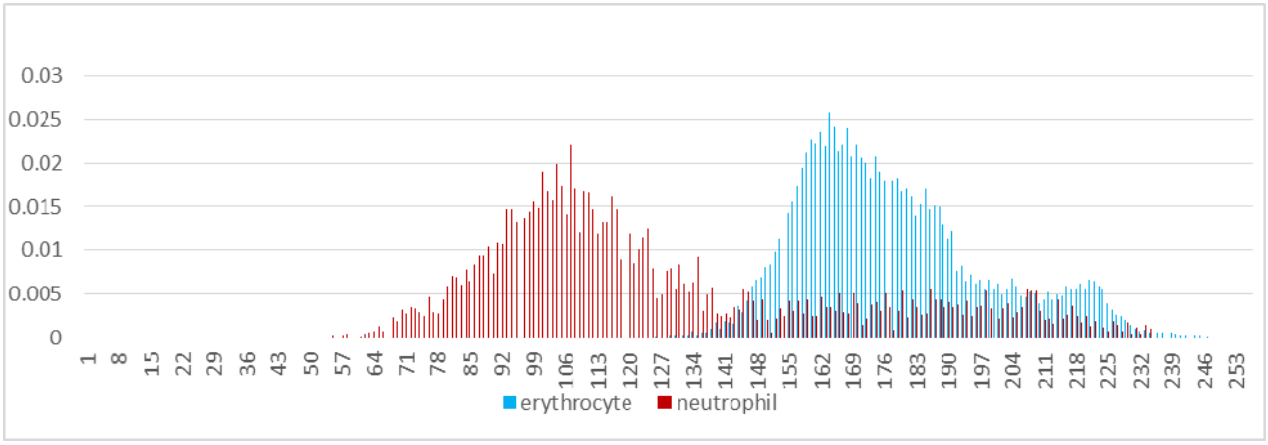

Fig. 4. Common intensity features

The distance between histograms can be calculated using Eq. 1.

$$D(a,b) = \sum_{i=0}^{n} |a_i - b_i| \quad (1)$$

Here, the $D(a, b)$ is the distance between histogram a and b; n – is the number of maximal possible values on the histogram (255 for intensity feature, 64 for color feature); a_i, b_i – is the i-th values of a and b histogram respectively.

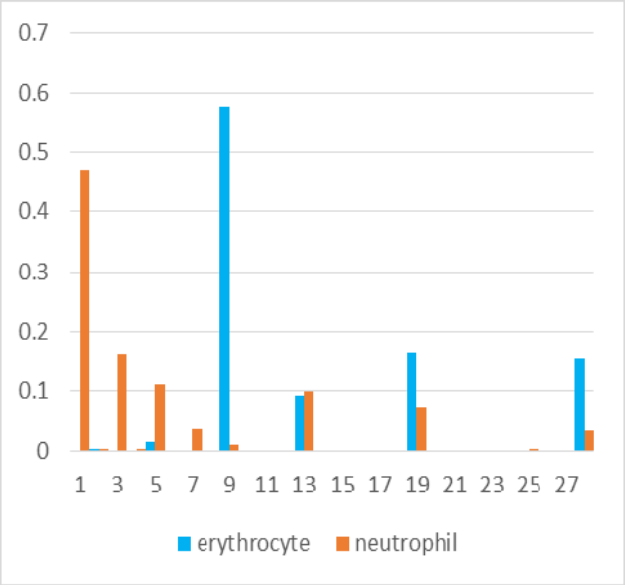

Fig. 5. Common color features

However, k-means clustering results can depends on selected initial centroids. Due to a series of experiments it was observed that inproper centroid detection may lead to incorrect classification results, when the cells that are realy the same type, may be classified as different types due to a little difference between them. For classifier training purpose we were force to improve initial centroid selection. We perform an algorithm that detects a mid histogram for a list of extracted histograms of specified cell type. Mid histograms are then considered as initial centroids for cluster analysis.

The algorithm that defines mid histograms looks as follows:

Pre: mid is the initially empty histogram which will contain resulting mid histogram after algorithm completes
set n **to** 0
for each histogram **in** list **do**
 for each key **in** histogram **do**
  set value **to** histogram.key
  set mid.key **to** ((mid.key * n) + histogram.key) / ++n
 end for
end for

VI. Cell Classification

Having common color or intensity histograms, we can perform classification of any random cell on the blood smear image. Histogram distance between selected cell feature and each cell type common feature can be applied to define the nearest appropriate cell class. Selecting the common feature with minimal distance to selected cell feature we will obtain proper cell type.

In case of multiple features we need to introduce similarity measure as a sum of normalized distances divided by the number of features. Distance normalization is required in this case because different features can be measured on different scales which can have different influence on similarity value. However, because we have histograms already normalized, the maximal possible distance betwee them equals one, so the similarity mesure can be written as in Eq. 2.

$$S(o,c) = \frac{\sum_{i=0}^{N} D(o_i,c_i)}{N} \quad (2)$$

Where $S(o, c)$ is the measure that define how similar is object o to class c; N is the number of features applied and $D(o_i,c_i)$ is the distance between i-th feature of object and i-th feature of class. In case of normalized histograms the similarity measure will always be in range [0;1]. This approach can be also used to find similar images in the images data base.

The results of blood smear image classification are illustrated in Fig. 6, where blue squares correspondes to erythrocytes, and red squares corresponds to neutrophils, which are the subtype of leukocytes.

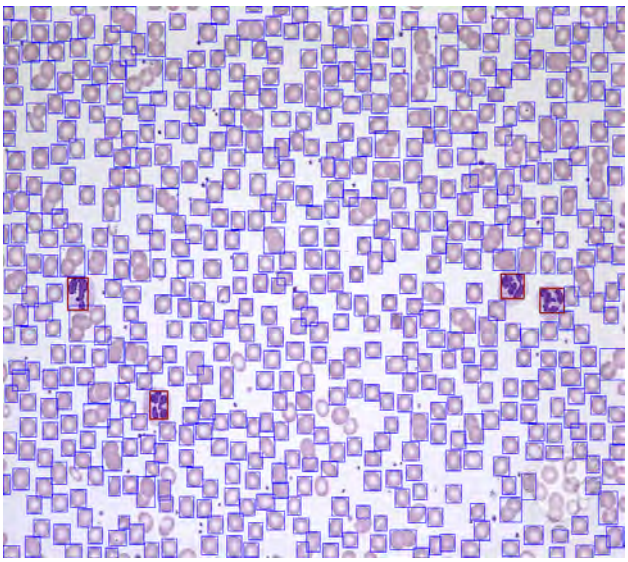

Fig. 6. Classified blood cells

VII. Results

Series of experiments with object detection were held to confirm the detection algorithm effectiveness. Due to Table 1, using a little negative threshold correction, 4-connectivity detection and removing components that are smaller than 65% size of average component, gives the best results of cell detection. This case is selected in row number five. Further decreasing of threshold and size filter coefficient leads to lower quality of cell detection.

TABLE 1

THE RESULTS OF CELL DETECTION EXPERIMENTS

No.	1	2	3	4	**5**	6	7
Tthreshold correction	0	-5	-15	-15	**-15**	+5	+10
Connectivity	8	8	8	4	**4**	4	4
Size filter	.75	.75	.75	.7	**.65**	.7	.7
Detected cells	537	540	550	578	**588**	542	525
Real cells	643	643	643	643	**643**	643	643
Precision (%)	83	84	85	90	**91**	84	82

TABLE 2

THE RESULTS OF CELL CLASSIFICATION EXPERIMENTS

No	1	2
Random initial centroids	+	-
Mid initial centroids	-	+
Erythrocytes (classified/real)	581/584	584/584
Neutrophils (classified /real)	4/4	4/4
Monocytes (classified /real)	0/0	0/0
Lymphocytes (classified /real)	3/0	0/0

Also, series of experiments with intensity and color feature clustering were held to confirm the effectiveness of classification approach and initial centroid detection algorithm. Due to Table 2, applying mid histograms as initial centroids for k-means cluster analysis improves the results of blood cells classification.

Conclusion

New approach for object detection and classification of blood smear images is proposed. Object detection is based on connected components labeling preceded by image binarization. To reduce the algorithm complexity segmentation was used. Influence of different connected component labeling, binarization threshold and size filter parameters on detection quality were demonstrated. The color and intensity features were used for objects comparison. Classifier was trained with the k-means feature clustering. It was proposed to use mid histograms as initial centroids for cluster analysis. Centroids obtained after cluster analysis were considered as common object class features. Minimal distance between extracted feature and common object class feature was considered to define object class. Software for classification of blood smear images was implemented. Some experiments with cells classification were held to confirm the algorithm effectiveness. It was noticed that it depends segmentation approach, labeling procedure, chosen features, initial centroid choosing approach.

References

[1] C. Hc sliding windows: Object localization by efficient subwindow search", CVPR, 2008.

[2] Pham, Dzung L.; Xu, Chenyang; Prince, Jerry L., "Current Methods in Medical Image Segmentation". Annual Review of Biomedical Engineering 2: 315–337, 2000.

[3] Luigi Di Stefano, Andrea Bulgarelli, "A Simple and Efficient Connected Components Labeling Algorithm," ICIAP, 10th International Conference on Image Analysis and Processing, pp.322, 1999.

[4] N. Otsu, ''A threshold selection method from gray level histograms,'' IEEE Trans. Syst. Man Cybern. SMC-9, 62–66, 1979.

[5] MacKay, David, "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999, 2003

[6] Orchard M, Bouman C, "Color quantization of images". IEEE Trans Signal Process 39(12):2677-2690, 1991.

[7] P. Maslak, "Normal peripheral blood smear - 1." http://imagebank.hematology.org/AssetDetail.aspx?AssetID=3666&AssetType=Asset, September 2008.