White Blood Cells Identification and Classification from Leukemic Blood Image

Lorenzo Putzu, Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, via Ospedale 72, 09124 Cagliari, Italy.

Abstract. The counting and classification of blood cells allows the evaluation and diagnosis of a vast number of diseases. Through the analysis of white blood cells (WBCs) the ALL - Acute Lymphocytic Leukemia, a blood cancer that can be fatal if left untreated, can be detected. Nowadays the morphological analysis of blood cells is performed manually by skilled operators. This involves numerous drawbacks, such as slowness of the analysis and a non-standard accuracy, dependent on the operator skills. In literature there are only few examples of automated systems in order to analyze and classify the blood cells, most of which only partial. This paper presents a complete and fully automatic method for WBCs identification and classification from microscopic images. The proposed method firstly individuates WBCs from which, subsequently, are extracted morphological features necessary for the final stage of classification. The whole work has been developed using MATLAB environment.

Keywords: Automatic detection, Classification, Image processing, Leukemia, Segmentation, White blood cell analysis.

1 Introduction

The counting and classification of blood cells from microscopic images allow evaluation and diagnosis of many diseases. Leukemia is a blood cancer that can be detected through the analysis of WBCs or leukocytes. Leukemia can be of two types: acute and chronic. According to the French-American-British (FAB) classification model, acute leukemia is classified into two subtypes: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In this paper we consider only the ALL, that affects a group of leukocytes called lymphocytes. The ALL primarily affects children and adults over 50 years and due to its rapid expansion into the bloodstream and vital organs can be fatal if left untreated [1]. Therefore, it becomes crucial early diagnosis of the disease for patients' recovery, especially in the case of children. The use of image processing techniques can help to count the cells in the human blood and, at the same time, provide information on the cells morphology. These techniques require only one image and are therefore less expensive, but at the same time more scrupulous in providing more accurate standards. The main goal of this work is the processing and analysis of microscopic images, in order to provide a fully automatic procedure to support the medical activity, able to count and classify the WBCs affected by ALL. Leukocytes are easily identifiable from microscopic images,

as their nuclei appear darker than the background, but data extraction from WBCs can present some complications due to wide variations in cell shape, dimensions and edges. The generic term leukocytes refers to a set of cells that are very different between them, which includes neutrophils, basophils, eosinophils, lymphocytes and monocytes, also distinguishable by the presence of granules in the cytoplasm and by the number of lobes in the nucleus. The lobes are the most substantial part of the nucleus and are connected to each other by thin filaments. Furthermore lymphocytes suffering from ALL, called lymphoblasts, have additional morphological changes that increase with increasing severity of the disease. In particular, lymphocytes present a regular shape, and a compact nucleus with regular and continuous edges. Instead, lymphoblasts present shape irregularities, small cavity in the cytoplasm, calls vacuoles, and spherical particles within the nucleus, called nucleoli. Therefore, in this paper, we present a method to identify all types of WBCs present in the microscopic images, which need various steps to reach the goal, and then classify WBCs as suffering from ALL or not. The identification of the leukocytes is carried out in the first step, described in Section 2. This step includes the identification and separation of grouped leukocytes and terminates with an image cleaning, through which all the abnormal components are removed from the image. The second step deal with the selection of the nucleus and the cytoplasm of each leukocyte, described in Section 3. The third step deals with the features extraction, described in Section 4, and the last phase proceeds to the classification of WBCs, described in Section 5. Each phase of the method, applied on a sample image, is analyzed in detail and compared with other approaches present in literature. The whole process can be schematized as showed in Figure 1.

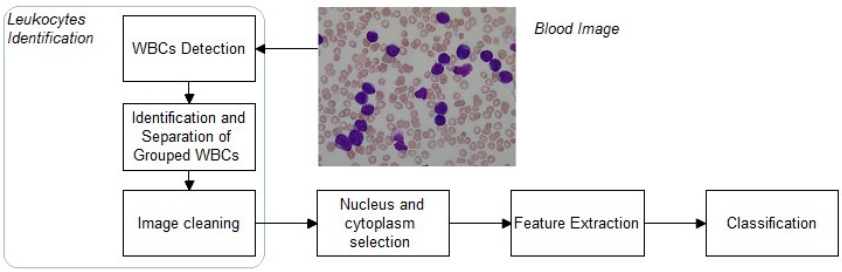

Fig. 1. Proposed method schema.

2 Leukocytes Identification

In many methods present in the literature the idea is to identify firstly the nuclei which are more prominent than other components [9] and then to detect the entire membrane, for example by region growing [2], [7]. In the proposed method instead, the membrane is detected firstly, in order to deal the subsequent separation of the adjacent cells more accurately. The WBCs identification was made possible thanks to the conversion in the CMYK color model. In fact, we have observed that leukocytes are more contrasted in the Y component of CMYK color model, this is because the yellow color is present in all the elements

of the image except in leukocytes, where it is practically absent (Fig. 2 shows an example). A redistribution of image gray levels is necessary in order to make easier the subsequent segmentation process. Then, an histogram equalization or a contrast stretching can be used at this stage. The segmentation is realized using a threshold automatically calculated. There are many threshold techniques available in literature [5]. Here, we use the threshold value based on triangle method or Zack algorithm [13] (see Fig. 3).

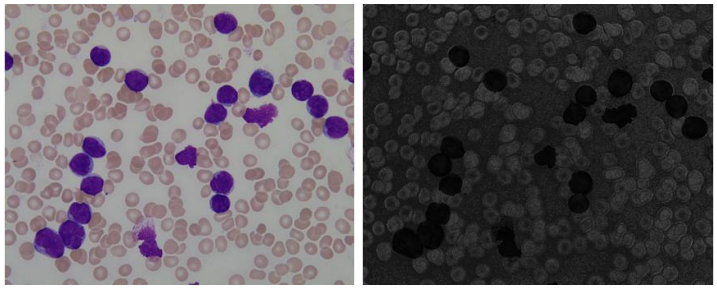

Fig. 2. Original RGB image and Y component image.

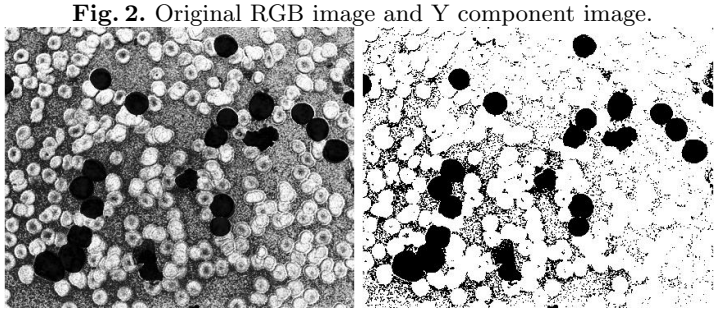

Fig. 3. Histogram equalization result and segmentation result.

To get a better result it is necessary to remove the image background. Some approaches for background extraction are present in literature, such as that showed by Scotti [12] that makes use of a collection of images for the estimation of the background pixels. The proposed approach involves the use of an automatic threshold to the original gray level image or along the green component of the RGB color space. The threshold value is calculated again using the triangle method (see Fig. 4). Background removal can be performed with simple arithmetical operations. In order to clean up the image we have used the operation called area opening, that allows to delete all the objects with a size smaller than the structuring element. The structuring element used has a circular shape and its size is calculated on the basis of the objects average in the image (see Fig. 5).

2.1 Identification and Separation of Grouped Leukocytes

Once obtained the image containing only the WBCs it is possible to verify if there are adjacent cells (or agglomerates of leukocytes) and, therefore, to provide for their separation. Several methods can be used to verify the presence of adjacent leukocytes [5]. In this work we have used the roundness value. Each connected

component having a low roundness value are classified as grouped leukocytes and so they must be separated. Some approaches to separate the adjacent cells, used by Kovalev [7], work on sub-images extracted from the original image by cutting a square around the nucleus previously segmented. So, assuming that each sub-image has a single WBC, a clustering around the nucleus is performed, by using the shape and the color information.

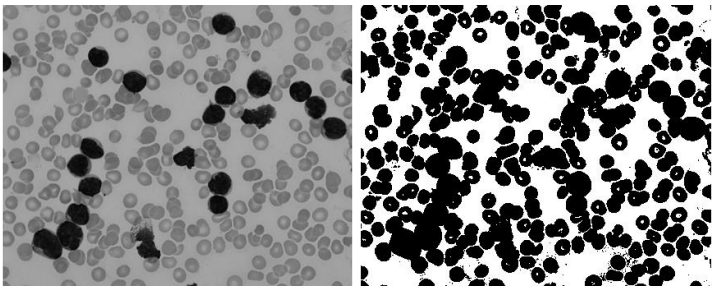

Fig. 4. Original gray level image and the segmentation result.

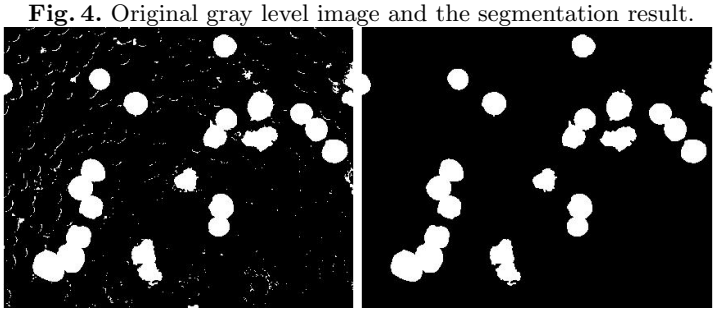

Fig. 5. Background removal result and area opening result.

The proposed approach is based on the method proposed by Lindblad [8] which uses the distance transform. The latter, applied to the binary image, associates to each pixel its distance from the border. A watershed segmentation is then applied to the distance transform to make a first separation between adjacent leukocytes. This approach performs well only in the presence of rounded leukocytes, but it does not perform equally well in the presence of multiple complex forms. For this reason it is necessary a second step to refine the contours extracted through watershed transform. Then, all the pixels of the component under examination which are located at a distance not greater than a predetermined value from the watershed line concerned, are taken in consideration. These pixels are then used to derive the deepest concavity for which the line of exact separation will have to pass. Therefore, by exploiting the information of the points of concavity and the information related to the points of maximum image in gray level, it is possible to obtain a cutting line that best fits the contour of the leukocytes, as it can be seen in Fig. 6.

2.2 Image Cleaning

The image cleaning requires the removal of all the leukocytes located on the edge of the image and of all abnormal components (that are not leukocytes), in

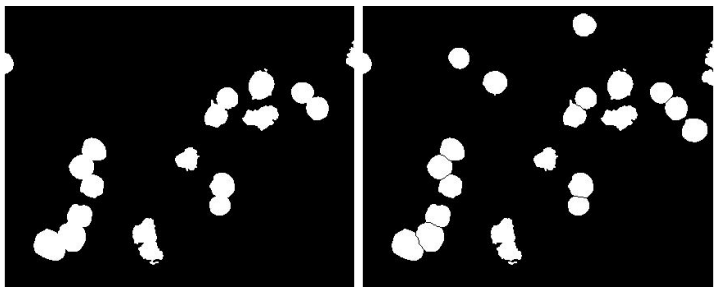

Fig. 6. Leukocytes identified as grouped and final separation results.

order to avoid errors in the later stages of the analysis process (see Fig. 7). The cleaning of the image edge is a simple operation, while the removal of abnormal components is a more complex process. To do this it is necessary to first determine the number of leukocytes present in the image. For each of them the size of the area and the size of the convex area are computed. The size of the area is used to calculate the mean area, necessary to determine and to eliminate the components with irregular dimensions. For example, a very small area might indicate the presence of artifacts not removed adequately, on the other side, a very large area may indicate the presence of adjacent leukocytes not separated adequately. Area and convex area are then used in combination for the calculation of the solidity value. All objects with a solidity value less than a predetermined threshold are discarded. In fact, even in this case, a value of solidity less than the threshold value indicates the presence of artifacts not removed adequately.

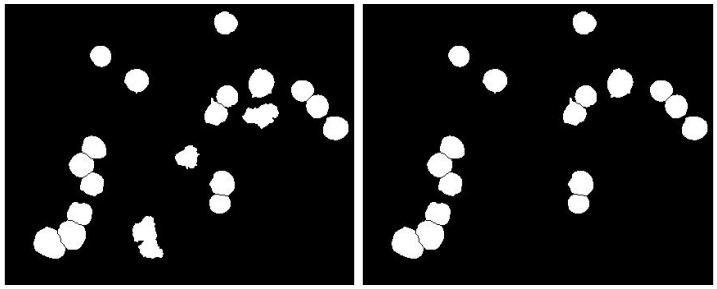

Fig. 7. Image after border cleaning and image after abnormal components removal.

3 Nucleus and Cytoplasm Selection

Once the leukocytes have been identified, it is possible to move to the second segmentation level that provides the selection of nucleus and cytoplasm. This step can be simplified performing an image crop using the bounding box size, that is the smallest rectangle that completely contains a connected component, with the aim to have a single leukocyte for each sub-image, as it is shown Fig. 8. A border cleaning operation is again necessary in order to preserve only the WBC in question. Since by definition, leukocytes nucleus is internal to the membrane, it is possible to perform further simplification, through the crop of the entire

portion of the image outside the leukocyte in question. This procedure allows a more robust nucleus selection, because it excludes artifacts from the selection completely.

Fig. 8. Gray level and binary sub-image of individual leukocytes. Border cleaning from binary image. Gray level sub-image with superimposed border. Gray level sub-image with external leukocyte image cropped.

Nucleus selection approach takes advantage from Cseke [3] observations, who found that WBCs nuclei are more in contrast on the green component of the RGB colour space. Threshold operation using Otsu [10] in this color space, however, does not produce clean results, especially with the presence of granulocytes, whose granules are selected erroneously as part of the nucleus. For this the binary image obtained from the green component is combined with the binary image, obtained from the a* component of the CIELab color space, again through a threshold operation. The mask obtained allows to extract clearly the leukocytes nucleus. At the end, to obtain the cytoplasm you just have to perform a subtraction operation between the binary image containing the whole leukocyte and the image containing only the nucleus (see Fig. 9).

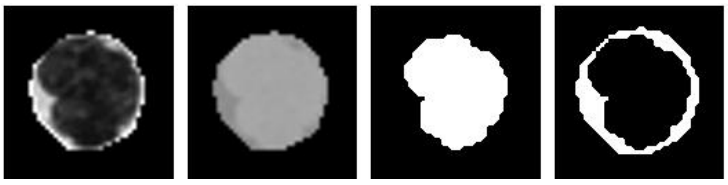

Fig. 9. Images of the G component in the RGB space and a* component in the CIELab space. Binary image after nucleus selection. Binary image after cytoplasm selection.

4 Feature Extraction

In this phase the idea is to extract the descriptors that best approach the visual patterns to which the pathologists refer and at the same time most relevant to the subsequent step of classification. Firstly, starting from the binary sub-images of cytoplasm and nucleus we have extracted shape descriptors such as area, perimeter, major axis, minor axis, orientation, eccentricity, rectangularity, compactness, convex hull, convex area, convex perimeter, convexity, roundness and solidity. At these classical measures are added two specific measures for the analysis of leukocytes, the ratio between the area of the cytoplasm and the nucleus and the number and structure of the lobes of the core. An approach proposed by Scotti [12] for the extraction of the number of nucleus' lobes makes use of repeated erosions until the correct number of lobes is reached. In a similar way our approach makes use of the ultimate erosion of the binary image [6],

that consists of the regional maxima of the Euclidean distance transform of the complement of the binary images. The result is shown in the Fig. 10, where it can be noted that, also in this way, the number of lobes remains unchanged.

Fig. 10. The binary image of the nucleus and the result of the extraction of the number of lobes obtained through iterative erosion and through ultimate erosion.

The gray level sub-images are instead used for the extraction of the color characteristics and texture characteristics. The color characteristics considered are mean, variance, standard deviation and entropy. While textural characteristics were calculated from the GLCM in order to characterize the leukocytes cytoplasm and are energy, entropy, contrast and correlation. These characteristics have been calculated for angles of 0, 45, 90 and 135 degrees. The total number of extracted features is then 50: 30 shape descriptors, 4 color descriptors and 16 texture descriptors.

5 Classification and Experimental Results

The proposed method was finally tested with the database ALL-IDB1 [4] which consists of 108 original blood sample images. Each image has an associated text file containing the coordinates of the centroid of each candidate linfoblast, manually estimated by a skilled operator. The test was carried out with a subset of 33 images acquired from the same camera and under the same lighting conditions. These images were taken with an Olympus C2500L camera and have a resolution of 1712x1368. From this sample of images, in the earlier stages of the analysis process, have been properly extracted 245 sub-images containing individual leukocytes, with an accuracy of 92%. The resulting dataset, composed of a matrix of features with size 50x245 and by a classification vector with size 1x245, can then be used to test the final stage of the process. The classification model chosen for this phase is the Support Vector Machine SVM, since this model is particularly suitable for binary classification problems, in which the separation between the classes depends on a large number of variables. The performance of the model were then evaluated by a 10-fold Cross-Validation. The numerical results are shown in Table 1.

Figures of Merit	Value
TP	184
TN	2
FP	58
FN	0
Missclassification	0,25
Sensitivity	1

Table 1. This table shows the performance of the classification model.

6 Conclusion

In this work we have proposed an innovative method for the completely automatic identification and classification of leukocytes by microscopic images, in order to provide an automated procedure as support medical activity, in recognition of acute lymphocytic leukemia. The results obtained show that the proposed method is able to identify in a robust way the WBCs present in the image, being able to properly classify all leukocytes suffering from disease and offering a good level of overall accuracy. Further developments of the proposed method could affect the separation step, which is of considerable importance in order to take into account all leukocytes in the image. In addition to increase the level of overall accuracy is required the use of a multi-class classification model for the identification of various types of leukocytes and finally of the lymphoblasts. It will also be necessary to expand the size of the dataset in order to provide to the classification model a greater number of useful examples in the training phase.

Acknowledgments. This work has been funded by Regione Autonoma della Sardegna (R.A.S.) Project CRP-17615 DENIS: Dataspace Enhancing Next Internet in Sardinia. We wish to thank Dr. Scotti for having made available the Dataset on which we could test our method.

References

1. Biondi, A., Cimino, G., Pieters,R., Pui, C. H., Biological and Therapeutic Aspects of Infant Leukemia, 2000
2. Cheewatanon, J., Leauhatong, T., Airpaiboon, S., and Sangwarasilp, M., A New White Blood Cell Segmentation Using Mean Shift Filter and Region Growing Algorithm, 2011
3. Cseke, I., A Fast Segmentation Scheme for White Blood Cell Images, 1992
4. Donida Labati, R., Piuri, V., Scotti, F., ALL-IDB: the Acute Lymphoblastic Leukemia Image DataBase for image processing, 2011
5. Gonzalez, R. C., Woods, R. E., Digital Image Processing, Prentice Hall Pearson Education, Inc.. New Jersey, USA, (2002)
6. Gonzalez, R. C., Woods, R. E., Eddins, S.L., Digital Image Processing Using MATLAB, Pearson Prentice Hall Pearson Education, Inc., New Jersey, USA, (2004)
7. Kovalev, V. A., Grigoriev, A. Y., Ahn, H., Robust Recognition of White Blood Cell Images, 1996
8. Lindblad, J., Development of Algorithms for Digital Image Cytometry, 2002
9. Madhloom, H. T., Kareem, S. A., Ariffin, H., Zaidan, A. A., Alanazi, H. O., Zaidan, B. B., An Automated White Blood Cell Nucleus Localization and Segmentation using Image Arithmetic and Automated Threshold, 2010
10. Otsu, N., A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66, (1979)
11. Scotti, F., Automatic Morphological Analysis for Acute Leukemia Identification in Peripheral Blood Microscope Images, 2005
12. Scotti, F., Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images, 2006
13. Zack, G., Rogers,W., Latt, S., Automatic Measurement of Sister Chromatid Exchange Frequency, 1977