Investigation of Different Classification Models to Determine the Presence of Leukemia in Peripheral Blood Image

Lorenzo Putzu and Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari,
via Ospedale 72, 09124 Cagliari, Italy
lorenzo.putzu@gmail.com, dirubert@unica.it

Abstract. The counting and classification of blood cells allows the evaluation and diagnosis of a vast number of diseases, such as the ALL - Acute Lymphocytic Leukemia, detected through the analysis of white blood cells (WBCs). Nowadays the morphological analysis of blood cells is performed manually by skilled operators, involving 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 able to process a whole image in order to analyze and classify all the WBCs included. This paper presents a complete and fully automatic method for WBCs identification from microscopic images and an evaluation of different classification model to determine the presence of leukemia. Experimental results show that the proposed method is able to identify the cells carrying leukemia and consequently to determine whether a patient is suffering from this disease.

Keywords: Automatic detection, Classification, Feature selection, Leukemia, Segmentation, White blood cell analysis.

1 Introduction

ALL is a blood cancer that influences a group of leukocytes called lymphocytes, and 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]. An early diagnosis of the disease is crucial for patients' recovery, especially in the case of children. The observation of blood samples under a microscope is one of the possible procedures for the diagnosis of ALL. This method suffers from slowness and provides a non-standard accuracy dependent on the operator skills. Image processing techniques can help to count the cells in the human blood quickly and, at the same time, provide more accurate information on the cells morphology. Unfortunately the generic term leukocytes refers to a set of cells that are very different between them, in shape and size, which includes neutrophils, basophils, eosinophils, lymphocytes and monocytes, so data extraction from WBCs can present some complications. Furthermore lymphocytes suffering from ALL, called lymphoblasts, have additional morphological changes, like

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 612–621, 2013.

shape and size irregularities, that increase with increasing severity of the disease. Therefore, in this paper we propose a fully automatic procedure to support the medical activity, able 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. The second step deals 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.

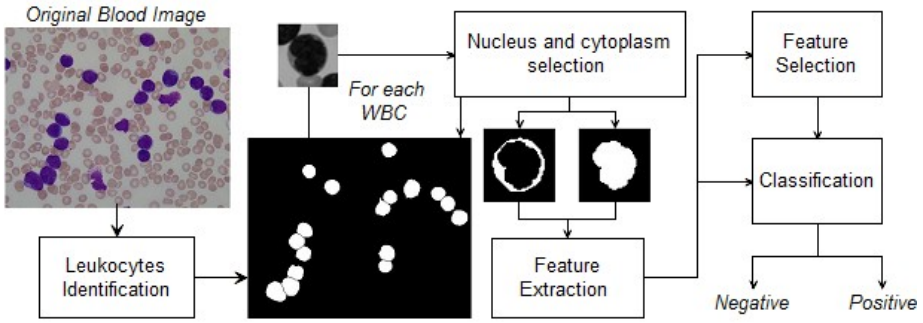


Fig. 1. Proposed method diagram

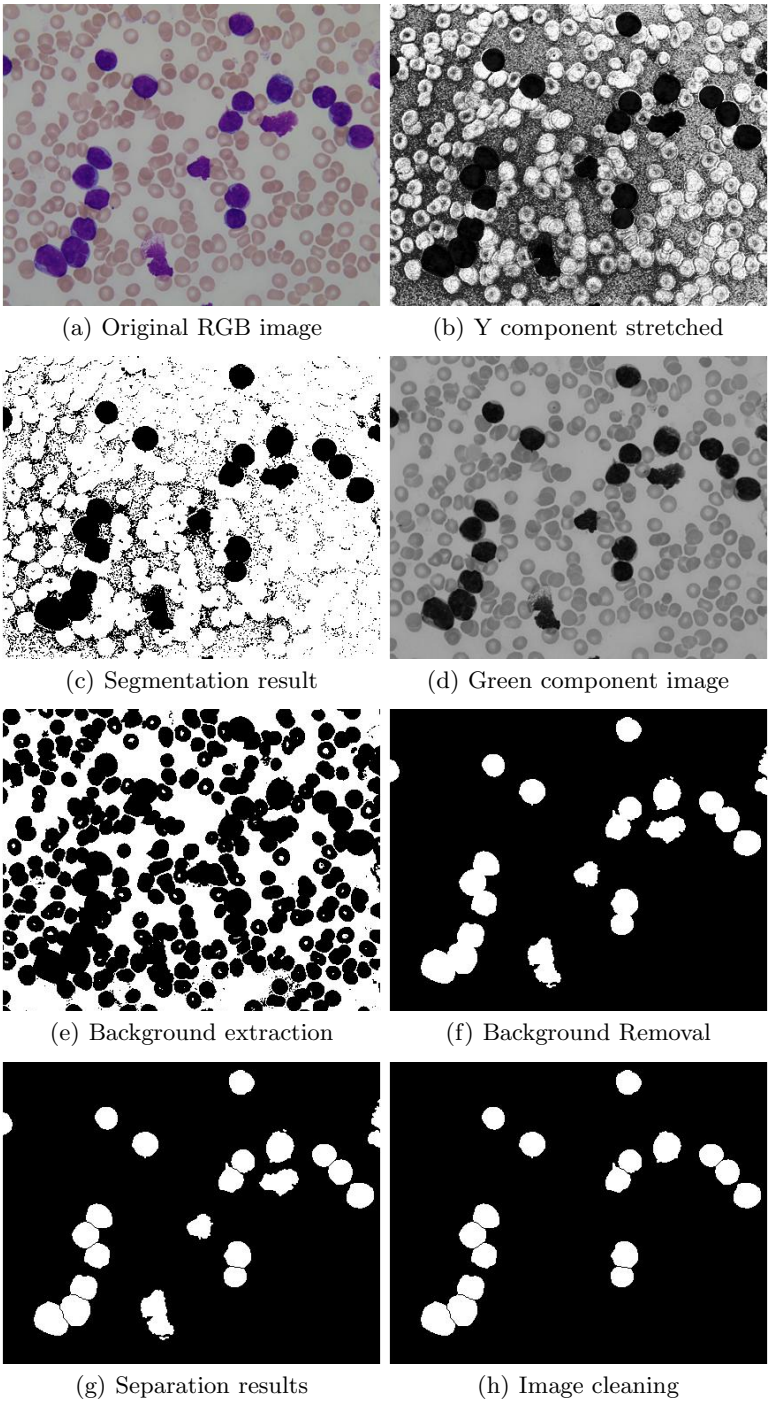
2 Leukocytes Identification

In many methods present in literature the idea is to identify firstly the nuclei which are more prominent than other components [8] and then to detect the entire membrane, for example by region growing [2], [6]. In the proposed method instead, the membrane is detected firstly thanks to the conversion in the CMYK color model, because the leukocytes are more contrasted in the Y component. After a redistribution of image gray levels by histogram equalization (Fig. 2(b) shows an example), we segment the image using a threshold automatically calculated through the triangle method or Zack algorithm [11] (see Fig. 2(c)), that we consider the best threshold technique available in the literature [5] for this application. 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 [10] 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, calculated again using the triangle method,

but this time starting from the G component of the RGB color space (see Fig. 2(e)). Background removal was performed with an arithmetical operation and an area opening operation (see Fig. 2(f)). Another problem to be addressed in analysis of blood image is the presence of agglomerates of leukocytes. Several methods can be used to verify the presence of adjacent leukocytes [5]. In our work each connected component having a roundness value lower than 0.80, identified during our experimentation as the optimal threshold, is classified as grouped leukocytes and so it must be separated. Some approaches to separate the adjacent cells, used by Kovalev [6], 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 shape and color information. Our approach works on the whole image, avoiding problems that may arise after nucleus identification (even the nuclei of white blood cells may be in contact) and is based on the method proposed by Lindblad [7] 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 along the border and for which passes a watershed line are considered as a concavity point, 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. 2(g). The last step for the identification of leukocytes consists of a cleaning process of the image, by removing the cells located along the borders and by removing all the components with irregular shape and size, as it can be seen in Fig. 2(h) (for details see [12]). The abnormal components are detected using a solidity value of 0.90, identified during the experiments performed as the most discriminatory value.

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 is performed from sub-images, created using the bounding box size and to which is applied an operation of border cleaning with the aim to have a single leukocyte for each sub-image, as it is shown Fig. 3(c). Since by definition, leukocytes nucleus is internal to the membrane, it is possible to crop the entire portion of the image outside the leukocyte in question, in order to excludes artifacts during nucleus selection. 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



(a) Original RGB image (b) Y component stretched

(c) Segmentation result (d) Green component image

(e) Background extraction (f) Background Removal

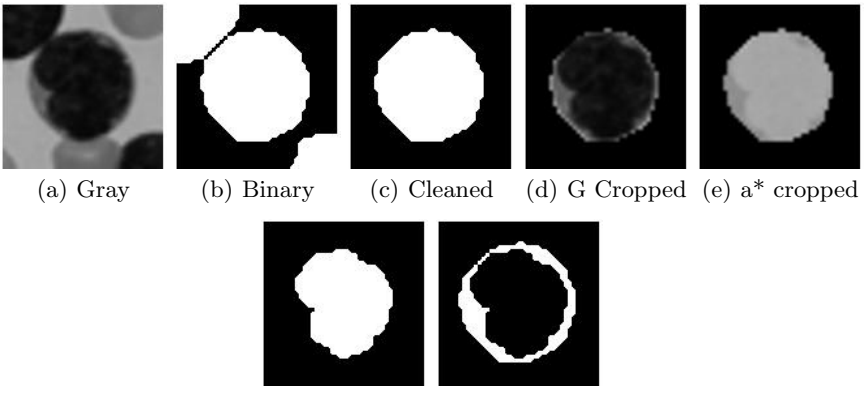
(g) Separation results (h) Image cleaning

Fig. 2. Leukocytes identification process

Otsu [9] 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. 3(f),3(g)).

4 Feature Extraction

Speak about feature extraction in this context means to transform the images into data, then extract information reflecting the visual patterns which the pathologists refer to, but at the same time it is necessary to extract the descriptors that are most relevant to the subsequent classification process. For this reason, from the sub-images calculated previously are extracted 3 different types of descriptors: shape features, color features and texture features. Starting from the binary sub-images of nucleus and cytoplasm 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. To these classical measures we added two specific measures for the analysis of leukocytes, the ratio between the area of the cytoplasm and the nucleus and the number of nucleus lobes (for details about lobes number extraction see [13]). The main disadvantage of the shape features is that they are very susceptible to errors in segmentation. For this reason, these descriptors were used together with regional descriptors less susceptible to errors. Among these there are the color descriptors, which are the most discriminatory



(a) Gray (b) Binary (c) Cleaned (d) G Cropped (e) a* cropped

(f) Nucleus (g) Cytoplasm

Fig. 3. Nucleus and cytoplasm selection process

features of the blood cells. The color descriptors used are mean, standard deviation, smoothness, skewness, kurtosis, uniformity and entropy, calculated from the sub-images in shades of gray. Often, however, the descriptors based only on histograms have some drawbacks as they do not give information on the mutual position of the pixels. Some objects have in fact a repeating pattern as the primary visual characteristic and so it is necessary to consider not only the intensity distribution but also the positions of the pixels having a similar gray level. Then we have evaluated the descriptors applied to the matrix of co-occurrence calculated starting from the sub-images in gray level. The descriptors are autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares (variance), sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation1, information measure of correlation2, inverse difference normalized and inverse difference moment normalized. These features have been calculated for angles of 0, 45, 90 and 135 degrees. The total number of extracted features is then 117: 30 shape descriptors, 7 color descriptors and 80 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% (detailed results can be observed in [12]). From these sub-images are then extracted a feature matrix with size 117x245 and a classification vector with size 1x245, which can be used to test the final stage of the process. In our previous work [13] the classification process was carried out using only 50 features and using the SVM classifier, since this model is particularly suitable for binary classification problems, in which the separation between the classes depends on a large number of variables. In this work, since the number of features is even higher we decided to proceed with different approaches. Firstly, the classification was performed using different models, both by exploiting the entire feature vector and by reducing its dimensionality through an operation of feature selection. The models tested for the classification process are still the SVM, K-Nearest Neighbor using different values of K, Naive Bayes by a Gaussian and Kernel data distribution, Decision Trees, Random Forest and different ensemble models such as AdaBoost, RobustBoost, LogitBoost, GentleBoost, Bag and Subspace. For the feature selection, firstly we compared different methods that use the sequential forward feature selection based on K Nearest Neighbor, Naive Bayes, Decision Trees, and Random Forest. In all cases,

Table 1. Experimental results with sequential forward feature selection: for each test accuracy (acc) and standard deviation (SD) are reported

	No FS		k-NN FS		NB FS		TREE FS		RF FS	
Classifier	acc	S D	acc	S D	acc	S D	acc	S D	acc	S D
SVM	0,76	0	0,80	0,009	0,85	0,004	0,87	0,009	0,89	0,006
NN k=1	0,73	0,016	0,90	0,005	0,89	0,003	0,74	0,006	0,83	0,004
NN k=2	0,73	0,015	0,90	0,007	0,81	0,009	0,74	0,006	0,83	0,005
NN k=3	0,75	0,014	0,91	0,005	0,85	0,004	0,79	0,009	0,84	0,003
NN k=4	0,75	0,011	0,91	0,004	0,85	0,007	0,79	0,011	0,84	0,007
NN k=5	0,73	0,011	0,89	0,006	0,85	0,007	0,79	0,008	0,84	0,007
NN k=6	0,74	0,011	0,89	0,007	0,85	0,006	0,79	0,006	0,85	0,003
NN k=7	0,73	0,009	0,88	0,006	0,86	0,007	0,79	0,009	0,85	0,003
NN k=8	0,75	0,014	0,88	0,005	0,86	0,008	0,79	0,009	0,86	0,006
NN k=9	0,72	0,007	0,87	0,006	0,85	0,004	0,82	0,006	0,83	0,005
NN k=10	0,73	0,021	0,88	0,004	0,85	0,004	0,81	0,008	0,84	0,004
NB	0,81	0,006	0,82	0,008	0,89	0,003	0,84	0,007	0,81	0,003
NBK	0,85	0,008	0,84	0,007	0,91	0,006	0,87	0,006	0,86	0,003
tree	0,87	0,016	0,82	0,015	0,86	0,016	0,88	0,014	0,89	0,005
RF	0,89	0,005	0,87	0,004	0,90	0,006	0,90	0,009	0,92	0,006
ADA	0,87	0,008	0,85	0,006	0,87	0,008	0,89	0,006	0,92	0,006
Robust	0,85	0,015	0,86	0,009	0,87	0,012	0,85	0,018	0,88	0,007
Logit	0,87	0,008	0,85	0,005	0,89	0,004	0,89	0,006	0,91	0,009
Gentle	0,87	0,005	0,85	0,005	0,88	0,010	0,89	0,011	0,92	0,008
Bag	0,90	0,006	0,87	0,003	0,90	0,003	0,90	0,012	0,92	0,004
Subspace	0,78	0,008	0,78	0,012	0,83	0,010	0,82	0,005	0,80	0,010
Mean	0,79	0,010	0,86	0,007	0,86	0,007	0,83	0,08	0,87	0,005

given the small size of the dataset used, the performance of the models were then evaluated by a 10-fold Cross-Validation. The experimental results are shown in Table 1.

The results obtained show that in general all the classification models benefit from the process of feature selection. In particular we can observe how the SVM accuracy increments more with more elaborate feature selection processes. It's interesting also to note that all classification models have better performance associated with feature selection based on the same classifiers, for example the k-NN improves more with feature selection based on k-NN, Naive Bayes improves with feature selection based on Naive Bayes and so on. The only classifiers that don't get substantial performance improvements are the ensemble classifiers, which by their nature are generally more robust.

The feature selection has been implemented also with the algorithm of ReliefF, which, unlike the methods of sequential feature selection, does not provides the best feature vector, but it returns a vector containing all the features sorted according to their relevance. The classification was then tested, using the same classification models previously seen, and by using an increasing number of features (from 5 to 25). Also in this case the performances of the models were then

Table 2. Experimental results with ReliefF algorithm: for each test accuracy (acc) and standard deviation (SD) are reported

	5 feat		10 feat		15 feat		20 feat		25 feat	
Classifier	acc	S D	acc	S D	acc	S D	acc	S D	acc	S D
SVM	0,82	0,008	0,85	0,001	0,85	0,002	0,78	0,006	0,8	0,004
NN k=1	0,75	0,006	0,8	0,009	0,87	0,007	0,74	0,002	0,78	0,013
NN k=2	0,75	0,004	0,82	0,008	0,87	0,011	0,74	0,02	0,79	0,002
NN k=3	0,77	0,01	0,83	0,007	0,89	0,009	0,74	0,008	0,79	0,015
NN k=4	0,78	0,004	0,84	0,007	0,9	0,007	0,73	0,007	0,8	0,017
NN k=5	0,78	0,003	0,83	0,008	0,9	0,007	0,73	0,006	0,82	0,012
NN k=6	0,78	0,011	0,84	0,003	0,9	0,004	0,73	0,016	0,83	0,006
NN k=7	0,79	0,011	0,84	0,003	0,9	0,007	0,74	0,009	0,82	0,006
NN k=8	0,80	0,006	0,84	0,007	0,9	0,007	0,74	0,005	0,84	0,011
NN k=9	0,80	0,01	0,83	0,004	0,89	0,005	0,74	0,015	0,82	0,009
NN k=10	0,80	0,008	0,83	0,006	0,9	0,006	0,75	0,008	0,84	0,009
NB	0,81	0,003	0,82	0,003	0,85	0,006	0,82	0,006	0,79	0,002
NBK	0,83	0,006	0,87	0,004	0,88	0,003	0,85	0,005	0,8	0,005
tree	0,80	0,019	0,85	0,007	0,85	0,008	0,77	0,004	0,8	0,002
RF	0,82	0,008	0,9	0,009	0,91	0,009	0,91	0,005	0,87	0,006
ADA	0,83	0,013	0,88	0,008	0,88	0,006	0,89	0,018	0,87	0,007
Robust	0,81	0,008	0,87	0,007	0,89	0,008	0,89	0,008	0,87	0,014
Logit	0,82	0,005	0,87	0,011	0,89	0,01	0,89	0,009	0,87	0,016
Gentle	0,81	0,01	0,87	0,011	0,88	0,016	0,88	0,016	0,87	0,016
Bag	0,83	0,012	0,91	0,005	0,90	0,01	0,91	0,013	0,88	0,012
Subspace	0,82	0,007	0,82	0,006	0,82	0,004	0,83	0,01	0,80	0,003
Mean	0,80	0,008	0,85	0,006	0,88	0,007	0,80	0,009	0,80	0,009

evaluated by a 10-fold Cross-Validation. The experimental results are shown in Table 2.

Even in this case the performances of classification models benefit from feature selection process, but unlike the previous case where the results of a single classification model were more or less good according to the feature selection technique used, in this case all classification models have a similar behavior. In fact for all of them the performance increases with the increase of the features taken into account, up to the maximum value obtained with a number of features equal to 15, then decreased again. From the data collected it was possible to trace the best features selected from various methods of feature selection during the test phase. These features are shown in Table 3, sorted by relevance.

To verify if the selected features are effectively the best we have tested again all classification models seen previously and each time using an increasing number of features from 3 to 15. The best results were obtained using the first 11 features, by which was obtained a value of average accuracy (for all classifiers) equal to 0.88. The value of maximum accuracy has been obtained also in this case with the Random Forest classifier with an accuracy equal to 0.92.

Table 3. Most used descriptors after feature selection

Relevance	Type	Feature
1	Shape	Nucleus Area
2	Shape	Area ratio
3	Texture	Inverse Difference
4	Shape	Minor Axis Length
5	Shape	Lobes number
6	Shape	WBC Area
7	Color	Skewness
8	Color	Kurtosis
9	Shape	WBC Compactness
10	Shape	Nucleus Convex Area
11	Texture	Autocorrelation
12	Shape	WBC Perimeter
13	Shape	WBC Convexity
14	Shape	WBC Convex Area
15	Shape	Nucleus Perimeter

6 Conclusions

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. So, starting from an original image we can process it entirely, without performing manual crops or without identifying a working area, and we are able to process it in a fully automatic way without need for manual intervention by the user. 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. We have also obtained a list of features which can be proposed as single set of features to be used for the classification of leukemic images. In fact with this set of features the level of accuracy is comparable to that one obtained with the best methods of feature selection and classification, but with the advantage of completely overthrowing the computation time that are dilated for the extraction of a large number of feature and for the features selection. Further developments could affect the classification phase, in fact to increase the level of overall accuracy it is required the use of a multiclass 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 and to allow us the use of a validation method different from 10-fold Cross-Validation.

Acknowledgments. This work has been funded by Regione Autonoma della Sardegna (R.A.S.) Project CRP-17615 DENIS: Dataspace Enhancing Next

Internet in Sardinia. Lorenzo Putzu gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.). 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. Blood 96(1), 24–33 (2000)
2. Cheewatanon, J., Leauhatong, T., Airpaiboon, S., Sangwarasilp, M.: A New White Blood Cell Segmentation Using Mean Shift Filter and Region Growing Algorithm. International Journal of Applied Biomedical Engineering 4, 30–35 (2011)
3. Cseke, I.: A Fast Segmentation Scheme for White Blood Cell Images. In: Proceedings of the IAPR International Conference on Image, Speech and Signal Analysis, vol. 3, pp. 530–533 (1992)
4. Donida Labati, R., Piuri, V., Scotti, F.: ALL-IDB: the Acute Lymphoblastic Leukemia Image DataBase for Image Processing. In: Proceedings of the ICIP International Conference on Image Processing, pp. 2045–2048 (2011)
5. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using MATLAB. Pearson Prentice Hall Pearson Education, Inc., New Jersey (2004)
6. Kovalev, V.A., Grigoriev, A.Y., Ahn, H.: Robust Recognition of White Blood Cell Images. In: Proceedings of the 13th International Conference on Pattern Recognition, pp. 371–375 (1996)
7. Lindblad, J.: Development of algorithms for digital image cytometry. Uppsala University, Faculty of Science and Technology (2002)
8. 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. Journal of Applied Sciences 10(11), 959–966 (2010)
9. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9(1), 62–66 (1979)
10. Scotti, F.: Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images. In: Proceedings of the IEEE Instrumentation and Measurement Technology Conference, pp. 43–48 (April 2006)
11. Zack, G., Rogers, W., Latt, S.: Automatic Measurement of Sister Chromatid Exchange Frequency. Journal of Histochemistry and Cytochemistry 25, 741–753 (1977)
12. Putzu, L., Di Ruberto, C.: White Blood Cells Identification and Counting from Microscopic Blood Images. In: Proceedings of the WASET International Conference on Bioinformatics, Computational Biology and Biomedical Engineering, vol. 73, pp. 268–275 (January 2013)
13. Putzu, L., Di Ruberto, C.: White Blood Cells Identification and Classification from Leukemic Blood Image. In: Proceedings of the IWBBIO International Work-Conference on Bioinformatics and Biomedical Engineering, pp. 99–106 (March 2013)