Fuzzy based Blood Image Segmentation for Automated Leukemia Detection

Subrajeet Mohapatra, Sushanta Shekhar Samanta, Dipti Patra and Sanghamitra Satpathi*
IPCV Lab National Institute of Technology Rourkela, Odisha, India
*Ispat General Hospital, Rourkela, Odisha, India
subrajeets@gmail.com

Abstract*—Acute lymphoblastic leukemia (ALL) are a group of hematological neoplasia of childhood which is characterized by a large number of lymphoid blasts in the blood stream. ALL makes around 80% of childhood leukemia and it mostly occur in the age group of 3-7. The nonspecific nature of the signs and symptoms of ALL often leads to wrong diagnosis. Diagnostic confusion is also posed due to imitation of similar signs by other disorders. Careful microscopic examination of stained blood smear or bone marrow aspirate is the only way to effective diagnosis of leukemia. Techniques such as fluorescence in situ hybridization (FISH), immunophenotyping, cytogenetic analysis and cytochemistry are also employed for specific leukemia detection. The need for automation of leukemia detection arises since the above specific tests are time consuming and costly. Morphological analysis of blood slides are influenced by factors such as hematologists experience and tiredness, resulting in non standardized reports. A low cost and efficient solution is to use image analysis for quantitative examination of stained blood microscopic images for leukemia detection. A fuzzy based two stage color segmentation strategy is employed for segregating leukocytes or white blood cells (WBC) from other blood components. Discriminative features i.e. nucleus shape, texture are used for final detection of leukemia. In the present paper two novel shape features i.e., Hausdorff Dimension and contour signature is implemented for classifying a lymphocytic cell nucleus. Support Vector Machine (SVM) is employed for classification. A total of 108 blood smear images were considered for feature extraction and final performance evaluation is validated with the results of a hematologist.

I. Introduction

Leukemia is a group of hematological neoplasia which usually affects blood, bone marrow, and lymph nodes. It is characterized by proliferation of abnormal white blood cells (leukocytes) in the bone marrow without responding to cell growth inhibitors [1]. This results in suppression of hematopoiesis and, thereafter, anemia, thrombocytopenia, and neutropenia. Immature WBC can also accumulate in various extramedullary sites, especially the meninges, gonads, thymus, liver, spleen, and lymph nodes. Hence due to excessive lymphoid blast or myeloid blast in the marrow they also flow into the peripheral blood stream [2]. Diagnosing leukemia is based on the fact that white cell count is increased with immature blast (lymphoid or myeloid) cells and decreased neutrophils and platelets. The presence of excess number of blast cells in peripheral blood is a significant symptom of leukemia. So hematologists routinely examine blood smear under microscope for proper identification and classification of blast cells [3]. Leukemia can be pathologically classified into acute and chronic on a broader sense.

In the present paper acute lymphocytic leukemia (ALL) is only considered and the objective is to classify a lymphocyte as a normal or a lymphoblast. Regardless of advanced techniques microscopic examination of blood slides still remains as a standard leukemia diagnosis technique. Hence microscopic examination is the most economical way for initial screening of leukemia patients. Manual examination of the slides are subjected to bias i.e. operator experience, tiredness etc resulting with inconsistent and subjective reports. So there is always a need for a cost effective and robust automated system for leukemia screening which can greatly improve the output without being influenced by operator fatigue.

Over years many automatic segmentation and leukemia detection methods for blood smear images have been proposed. Broadly most of the methods are based on local image information. A two step segmentation process using HSV color model is used in [4]. Cell segmentation using active contour models is presented in [5]. There are several similar researches on blood cell segmentation and detection in the literature. Due to complex nature of the blood smear images and variation in slide preparation techniques much work has to be done to meet real clinical demands. It was concluded from the literature survey that the automation process solely depends upon proper segmentation and feature extraction. In the present paper we propose a method to automate peripheral blood film examination which can supplement the physician with objective data for better diagnosis and treatment. The technique we propose first separates the leukocytes from the other blood cells and then extracts the lymphocytes from the subclass. Fractal features, shape features and other texture features are extracted from those lymphocytes. Two new features for cell nucleus boundary roughness measurement is proposed here for leukemia detection. Support vector machine (SVM) is employed for classifying the nucleus images based on the extracted features into healthy and leukemic. Rest of the paper is organized as follows: Section II describes the framework of the proposed method. Experimental results are presented in Section III and Section IV presents a detailed analysis on the results obtained. Finally, Section V provides the concluding remarks.

978-1-4244-9190-2/11/$26.00 ©2011 IEEE

II. METHODS

The procedure for leukocyte classification in microscopic images consists of preprocessing, segmentation, feature extraction and classification. The overall working principle is depicted in Fig. 1. The blood smear image consists of red blood cells (RBC), white blood cells (WBC), and platelets. The proposed method is based on color image segmentation and our objective is to separate WBC from the background and finally separate nucleus and cytoplasm. As per acute leukemia is concerned the cytoplasm is scanty so we have considered only the nucleus as the region of interest and its essential features are extracted.

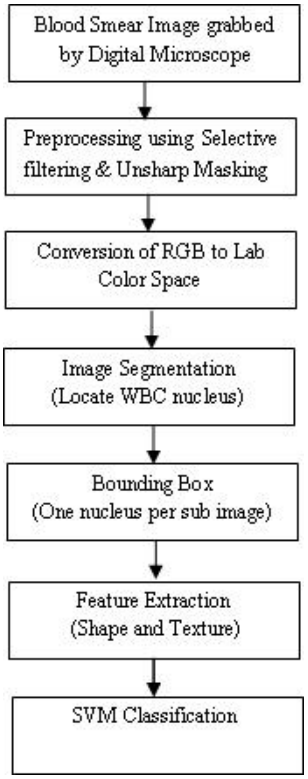


Fig. 1. System Overview

A. Image Grabbing

Leishman stained blood slides are collected from Ispat General Hospital, Rourkela Orissa. The images were captured with a digital microscope (Carl Zeiss India) under 100X oil immersed setting and with an effective magnification of 1000. Few images with permission from University of Virginia were also taken and were used for experimental purposes.

B. Preprocessing

Noise may be accumulated during image acquisition and due to excessive staining. All the test images are subjected to selective median filtering followed by unsharp masking [6]. Incorporation of adaptive threshold into the noise detection process led to more reliable and more efficient detection of noise.

C. Color Conversion

Typically images generated by digital microscopes are usually in RGB color space which is difficult to segment. Whereas the L*a*b* color space is an color representation technique which is basically used to reduce the color dimension from three to two in comparison to RGB. The L*a*b* space consists of a luminosity layer L*, chromaticity layer a* and chromaticity layer b*. Here the color information is represented in two components i.e. a* and b*. Due to less color dimension L*a*b* color space is mostly employed in color based clustering. In the present work microscopic images are converted from RGB color space to L*a*b* before clustering.

D. Image Segmentation

Segmentation is performed in two stages for extracting WBC nucleus from the blood microscopic images using color based clustering. Initial segmentation is achieved by an improved version of fuzzy clustering technique viz. Gustafson Kessel clustering [7], followed by nearest neighbor classification in L*a*b* space. Gustafson Kessel (GK) is a semi supervised clustering technique which is used to create K clusters from n observations. GK is an variation of Fuzzy C Means (FCM) which was developed in in 1973 by Dunn and improved by Bezdek in 1981. It attempts to achieve partition such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible [8]. Each pixel of an object is classified into four clusters based on orresponding a* and b* values in L*a*b* color space. The four clusters represents four regions i.e. RBC, WBC nucleus, cytoplasm and background stain. It was observed that WBC cytoplasm and RBC are classified into same cluster. In order to overcome the undesirable overlapping of regions, a second stage segmentation is performed using nearest neighbor classification. In the second stage we select a sample region randomly from each of the four clusters obtained using GK clustering. The mean color of the each sample regions are calculated in a*b* space and those values act as color indicators. Now each pixel in the L*a*b* space is classified into any of the four classes by calculating the euclidean distance between that pixel and each color indicator. Each pixel of the entire image will be labeled to a particular color depending on the minimum distance from each indicator. The nucleus segmented RGB image is reconstructed from the labeled image. We have only considered the cluster which contains blue nucleus as it is required for feature extraction and hence leukemia detection. Few left out holes in the nucleus creates problem during texture extraction and hence they are filled using morphological reconstruction.

E. Sub Imaging

As peripheral blood smear images are relatively larger usually, the cluster images are also large. But for accurate leukemia detection each nucleus feature has to be extracted individually for classifying it as a blast cell. Sub images containing single nucleus per image are essential for feature

extraction and were obtained using bounding box [9] technique. Using image morphology we select only those sub images which contain lymphocytes. The nucleus sub images of neutrophils, eosinophils and basophils are not considered for feature extraction as they are not associated with lymphocytic leukemia.

F. Feature Extraction

Feature extraction in image processing is a technique of redefining a large set of redundant data into a set of features (or feature vector) of reduced dimension. This transformation of the input data into the set of features is called feature extraction [10]. In the present paper broadly three types of features are extracted i.e. fractal dimension, shape features including contour signature and texture. In addition color features are also extracted from the nucleus image.

1) Fractal Dimension: Fractals have been used in medicine and science earlier for various quantitative measurement [11] [12]. Perimeter roughness of nucleus is a important measure that decides whether a particular nucleus represents a lymphoblast or a mature lymphocyte. Fractal geometry is a more convenient way to parameterize the cell boundary surface in comparison to euclidean geometry. Hausdorff dimension is an essential feature for fractal geometry and will be an essential quantitative measure for cell boundary roughness measurement. The procedure for Hausdorff Dimension measurement using box counting method [13] is introduced below as an algorithm:

1. Each nucleus color (RGB) image is converted to gray and successively to binary image.
2. Nucleus edge boundary is extracted using Canny [9] edge detection technique.
3. A grid of N squares is superimposed over the edges, while counting the edge occupied squares.
4. Step 3 is continued for an increasing number of squares.
5. The Hausdorff Dimension (HD) may then be defined as in (1).

$$HD = \frac{\log(N)}{\log(N(s))} \tag{1}$$

where, N is the number of squares in the superimposed grid and $N(s)$ is the number of occupied squares or boxes (box count). Higher HD signifies higher degree of roughness.

2) Contour Signature: Ill-defined or rough boundary is a significant feature for labeling a WBC nucleus as a blast cell. Along with the fractals contour signature method is also followed to measure the irregularity quantitatively. The nucleus boundary can be represented by a contour of dimension two. A better way of irregularity measurement of the contour is converting from coordinate based representation to distances from each contour point or edge pixels to a reference point. Since most nucleus have irregular shapes a convenient reference for the entire contour is the centroid or centre of mass. Euclidean distance measurement from the centroid to the contour points is described as follows:

1. Nucleus boundary pixel indices are obtained from the edge image which is obtained during HD measurement.
3. Centroid of the nucleus region is calculated using the edge pixels which represents a contour ((2)).
4. Euclidean distance is calculated from each boundary pixel to the centroid.
5. To measure the irregularity of the nucleus boundary variance (σ^2) of all the distances from the centroid obtained in step 4 is calculated.

$$\bar{x} = \frac{1}{M}\sum_{n=0}^{M-1} x(\mathrm{n}), \bar{y} = \frac{1}{M}\sum_{n=0}^{M-1} y(\mathrm{n}) \tag{2}$$

where (x,y) are the coordinates of the pixels along the contour and N is the total no of pixels on the contour.

3) Shape Features: According to hematologist the shape of the nucleus is an essential feature for discrimination of blasts. Region and boundary based shape features are extracted for shape analysis of the nucleus. All the features are extracted from the binary equivalent image of the nucleus with none zero pixels representing the nucleus region. The quantitative evaluation of each nucleus is done using the extracted features under two classes i.e. region based and boundary based. The features are as follows:

- *Area*: The area was determined by counting the total number of nonzero pixels within the image region.
- *Perimeter*: It was measured by calculating distance between succesive boundary pixels.
- *Compactness* : Compactness or roundedness is the measure of a nucleus as defined in (3).

$$Compactness = \frac{\mathrm{Perimeter}^2}{Area} \tag{3}$$

- *Solidity* : The ratio of actual area and convex hull area is known as solidity and is also an essential feature for blast cell classification. This measure is defined in (4).

$$Solidity = \frac{\mathrm{Area}}{ConvexArea} \tag{4}$$

- *Eccentricity* : This parameter is used to measure how much a shape of a nucleus deviates from being circular. It's an important feature since lymphocytes are more circular than the blast. To measure this a relation is defined in (5).

$$Eccentricity = \frac{\sqrt{\mathrm{a}^2 - b^2}}{a} \tag{5}$$

where a is the major axis and b is the minor axis of the equivalent ellipse representing the nucleus region.

- *Elongation* : Abnormal bulging of the nucleus is also an feature which signifies towards leukemia. Hence nucleus bulging is measured in terms of a ratio called elongation. This is defined as the ratio between maximum distance (R_{max}) and minimum distance (R_{min}) from the center of gravity to the nucleus boundary and is given by (6).

$$Elongation = \frac{\mathrm{R_{max}}}{R_{min}} \tag{6}$$

- *Formfactor* : This is an dimensionless parameter which changes with surface irregularities and is defined as (7).

$$Formfactor = \frac{4 \times pi \times Area}{\text{Perimeter}^2} \quad (7)$$

G. Color Feature Extraction

Since color is an important feature that human perceive while visualizing it is considered for extraction from nucleus regions. Hence for each nucleus image the mean color values in RGB and HSV color spaces are obtained.

H. Texture Features

Nucleus texture measurements were performed on gray scale version of the nucleus images. These features were computed from the co-occurrence matrices for each nucleus image. This includes

- *Homogeneity* : It is a measure of degree of variation.
- *Energy* :Is used to measure uniformity.
- *Correlation* : This represents correlation between pixel values and its neighborhood.
- *Entropy* : Usually used to measure the randomness.

I. Classification

Classification is the task of assigning to the unknown test vector, a label from one of the known classes. Since the patterns are very close in the feature space, support vector machines (SVM) are employed for classification. SVM is a powerful tool for data classification based on hyper plane classifier [14]. This classification is achieved by a separating surface (linear or non linear) in the input space of the data set. They are basically two class classifiers that optimize the margin between the classes [15]. The classifier training algorithm is a procedure to find the support vectors. Relevant extracted features as described in Section II-F are used as input to the SVM.

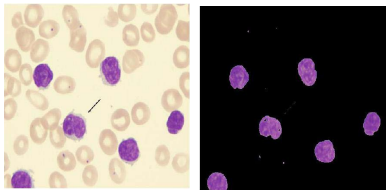
III. Experimental Results

The proposed technique has been applied on 108 peripheral blood smear images obtained from two places as mentioned earlier. The superiority of the scheme is demonstrated with the help of an experiment.

A. Experiment

A microscopic blood image of size 512×512 (Fig. 2(a)) is considered for evaluation. The input image is processed sequentially as per the steps mentioned in Section II. The segmented output of cell nucleus image obtained after applying K-means clustering algorithm is shown in Fig. 2(b). The cluster image containing only blue nucleus is used to obtain the sub images containing a single nucleus each as shown in Fig. 3.

Feature extraction can be done using methods as presented in Section II-F over each nucleus. Initially we use fractal geometry i.e. Hausdorff Dimension for calculating the perimeter roughness of each nucleus using the procedure explained in Section II-F1. Fig. 4(a) shows the nucleus boundary whose



(a) Input Image (b) Segmented Output

Fig. 2. Two Stage Segmentation Results

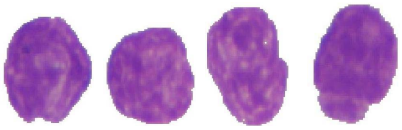


Fig. 3. Separeted Nucleus Sub Images using Bounding box technique

roughness is measured by Hausdorff Dimension using box counting method. A graphical plot showing box counting algorithm results is presented in Fig. 5. The straight line in the plot represents a line of best fit. And HD is obtained from the polynomial coefficients of the line of best fit. The $HD$is found to be 1.033 for the nucleus image shown in Fig. 4(a).

Fig. 4. Nucleus Boundary Detection. (a) Nucleus Image (b) Edge Image

The centroid of the nucleus contour is determined using the relation 2 . Euclidean distance between the centroid and boundary pixels is measured as shown in Fig. 5. The variance of all the distances is found to be 3.6355. Similarly the shape features are also measured using the relations given in Section II-F3. Few shape feature comparison between a mature lymphocyte and a lymphoblast is tabulated in Table I.

Color and texture features are also extracted for the image nucleus sample and recorded. Few texture measurements are tabulated in Table II. Among all the features the most relevant features are selected and used to train the SVM.

IV. Analysis

A set of quantitative features obtained from two known sample nucleus images is presented here. Our goal is to obtain a new set of features for better classification. We have tried to identify such features which are basically followed by hematologists. The results obtained in terms of features are also verified by an expert. The advantage of the proposed

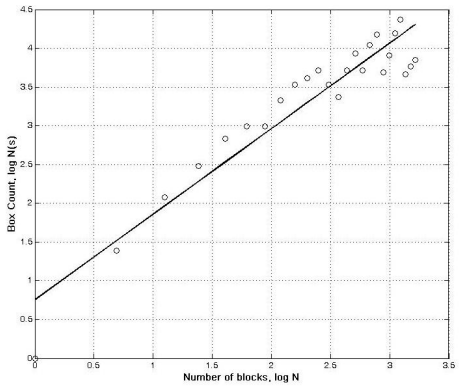


Fig. 5. Box Counting Algorithm Results

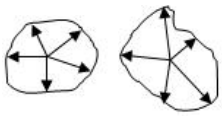


Fig. 6. Contour Signature

TABLE I
RESULTS OF VARIOUS SHAPE MEASUREMENTS

Measure	Lymphocyte	Lymphoblast
Area	2208	2715
Perimeter	186.1	208.3
Compactness	15.7	16
Solidity	1.0	1.0
Eccentricity	0.5	0.4
Elongation	1.4113	1.4181
Formfactor	0.2549	0.2504

TABLE II
RESULTS OF VARIOUS TEXTURE MEASUREMENTS

Measure	Lymphocyte	Lymphoblast
Homogeneity	0.877	0.879
Energy	0.277	0.292
Correlation	0.832	0.779
Entropy	2.787	3.305

scheme over existing schemes is that we are considering smear images with many lymphocytes. Existing schemes mostly consider those images which only have one lymphocyte under the field of view. Images from microscope are usually difficult and always require human intervention which is not desirable in an automated system. The proposed scheme using the new features is certainly aiming one step ahead towards automated system. The features extracted in our proposed scheme from the available images were used for training SVM classifier and an accuracy of 93% was observed.

V. CONCLUSION AND FUTURE WORK

A two stage WBC nucleus segmentation of stained blood smear images followed by relevant feature extraction for leukemia detection is the main theme of the paper. The paper mostly concentrates on measuring nucleus boundary irregularities using two methods i.e. hausdorff dimension and contour signature. Along with this shape, color and texture features are also considered for better detection accuracy. Leukemia detection with the proposed features were classified with SVM classifier. Results obtained encourage future works which includes classification of lymphoblast into various sub-types. Alternate techniques can be investigated for touching cells, stain independent image segmentation and leukemia type classification.

VI. ACKNOWLEDGEMENT

We are grateful to Charles E. Hess, M.D. and Lindsey Krstic, B.A. University of Virginia for providing microscopic images. The authors would like to thank Dr. S. Satpathy of IGH Rourkela for providing ground truth information and guidance which helped us in significantly improving the paper.

REFERENCES

[1] B. J. Bain . *A Beginner's Guide to Blood Cells*. Blackwell Publishing 2nd edition, 2004.
[2] Childrens Hospital of Wisconsin Website. http://www.chw.org
[3] C. Haworth, A. Hepplestone, P. Morris Jones, R. Campbell, D. Evans, M. Palmer. Routine Bone Marrow Examination in the Management of Acute Lymphoblastic Leukaemia of childhood. . *Journal of Clinical Pathology*, 34: 483 – 485, 1981.
[4] N. Sinha and A. G. Ramakrishnan. Automation of Differential Blood Count. *In Proceedings Conference on Convergent Technologies for Asia-Pacific Region*, 2:547 – 551, 2003.
[5] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, and S. Beksak, An Automated Differential Blood Count System.. In *Int. Conf. of the IEEE Engineering in Medicine and Biology Society* , volume 3, pages 2583 - 2586, 2001.
[6] Subrajeet Mohapatra Development of Impulse Noise Detection Schemes for Selective Filtering Master Thesis, National Institute of Technology Rourkela, 2008
[7] Daniel Graves and Witold Pedrycz Fuzzy C-Means, Gustafson Kessel FCM, and Kernel Based FCM: A Comparative Study Analysis and Design of of Intelligent Systems using Soft Computing Techniques , Springer, 2007.
[8] K. S. Ravichandran and B. Ananthi. Color Skin Segmentation using K–Means Cluster. *International Journal of Computational and Applied Mathematics*, 4(2):153 – 157, 2009.
[9] A. K. Jain, *Fundamentals of Digital Image Processing*. Pearson Education, 1st Indian edition, 2003.
[10] The Wikipedia the Free Encylopedia Website. http://en.wikipedia.org
[11] B. B. Mandelbrot. How long is the coast of Britain? Statistical self similarity and fractional dimension. *Science*, 156:636 – 638, 1967.
[12] B. T. Milne. Measuring the fractal geometry of landscapes. *Applied Mathematics and Computation*, 27:67 – 79, 1988.
[13] A. P. Pentland. Fractal based description of natural scenes . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:661 – 674, 1984.
[14] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.
[15] M. Hearst. Support vector machines. *IEEE Transactions on Intelligent Systems*, 18 – 28, 1998.