Blood Cell Segmentation using EM algorithm

Neelam Sinha
Bio Lab, Department of EE
Indian Institute of Science, Bangalore.
neelam@ee.iisc.ernet.in

A.G.Ramakrishnan
Bio Lab, Department of EE
Indian Institute of Science, Bangalore.
ramkiag@ee.iisc.ernet.in

Abstract

We present a method for segmentation of cells from color images of blood smears in the frame work of statistical modeling. The 2-part approach results in locating each of the white blood cells (WBC) and identifying the regions corresponding to the nucleus and the cytoplasm, in the given blood smear. The given RGB image is first converted to its Hue (H), Saturation (S), Value (V) equivalent. Each pixel is treated as a vector of the three dimensions namely H, S and V. The components are weighted to give more importance to the most distinguishing features. We segment by modeling each of the above mentioned regions by a distinct 3-D Gaussian distribution. In the first step, K-means clustering is performed on the 3-D feature vectors. This results in partitioning of the image into distinct regions. The centroids and the variances obtained in the K-means step are used to initialize Gaussian parameters for Expectation-Maximization (EM) algorithm. The EM algorithm iterates between segmentation and parameter estimation till convergence. A total of 115 images of smears were analyzed using our algorithm and successful segmentation was achieved in 80% of the cells contained in the images. The most important feature of this technique is that there are no parameters to be tuned by the user.

Keywords: Color image processing, blood cells, segmentation, EM algorithm, Clustering.

1 Introduction

To automate analysis of Leukaemic diseases, automated blood cell segmentation needs to be accomplished. A typical blood smear consists of white blood cells (WBC), red blood cells (RBC), plasma and platelets. The goal of segmentation is to locate the WBCs and to mark their nucleus and cytoplasm regions. This will facilitate their further processing to classify them as belonging to a particular class, or declaring them to be either healthy or diseased. The accuracy of segmentation is crucial since the subsequent steps in the analysis depend on it. Numerous segmentation methods have been proposed for digitized cell images of peripheral blood or bone marrow smears.

Dorin Comaniciu *et al.* [1] use non-Gaussian clusters in LUV color space. Their cell segmentation algorithm detects clusters in the L U V color space and delineates their borders by employing the gradient ascent mean shift procedure. Park [2] has carried out segmentation using Watershed algorithm. The nuclei of the WBCs are identified based on their size. This is followed by snake algorithm in order to draw the cell boundary. Methodical thresholding of the histogram is used to eliminate the background. A technique based on edge detection is proposed by Ravi *et al.* [6]. Here, the nucleus is segmented based on the edges that are effectively detected by Teager Energy operator proposed by Kaiser. Cytoplasm is segmented using selective mathematical morphology. Katz [5] suggests extraction of the region of interest from a larger image around thresholded cell nuclei. The segmentation of that image into cell and non-cell regions is carried out using Canny edge detection followed by a circle identification algorithm. Wermser *et al.* [4] have introduced a hierarchical thresholding scheme using a priori information regarding chromatic properties of background and cell components. Kovalev *et al.* [9] have proposed a three- step algorithm to segment white blood cells, employing prior knowledge of color information and using a circle-shaped approximation, Cseke [3] investigated the multi-step segmentation scheme, which implements the automatic thresholding method suggested by Otsu [7].

The performance of any of the above segmentation techniques will be limited by one or more of the following factors: significant case-specific distinctions in blood smear preparation, smear staining and image acquisition conditions. Further, most techniques mentioned here are sensitive to the right selection of parameters such as, threshold, mask-size and initial contour. Also, the assumption of circular shape is untenable in the case of most of the abnormal cells. Hence, we devise a robust technique free from the above assumptions and the need for user-interaction to tune parameters. In this paper, we report that two-part segmentation scheme that enables us to distinguish the WBC-cytoplasm and nucleus from the input image of a blood smear.

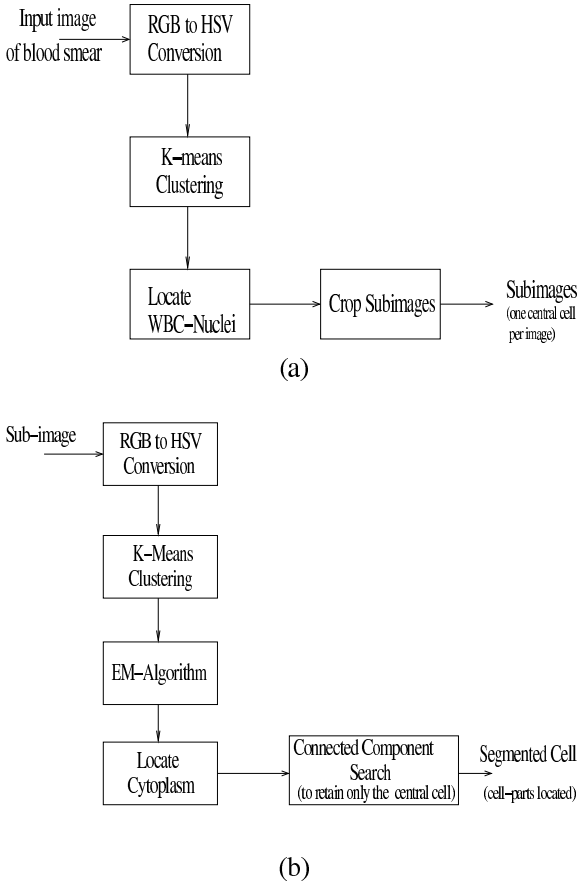


Figure 1: System Overview (a) Stage 1 (b) Stage 2

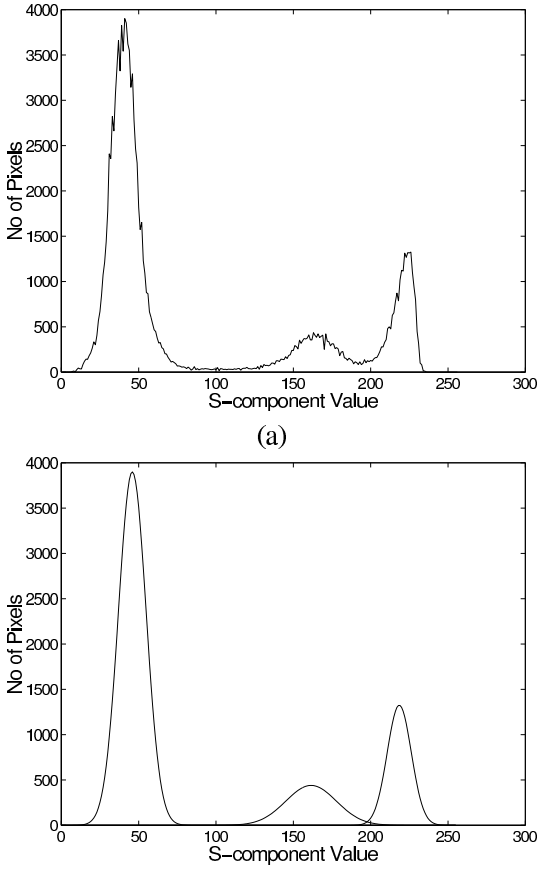


Figure 2: (a) Original histogram (b) corresponding Gaussian fit

2 Segmentation scheme

Fig. 1 shows the schematic of the proposed segmentation scheme. Our approach to segmentation is color-based. We first locate the nuclei of the cells using K-means clustering on the HSV equivalent of the image. We then crop a rectangular region around it that encompasses the entire cell. This is shown in Fig. 1a. Subsequent processing is carried out on the HSV equivalent of these sub-images. K-means clustering, followed by EM-algorithm are used to get the final segmentation of the cytoplasm and the nucleus regions. Protrusion of neighboring cells is removed using Connected Component Analysis. This is shown in Fig. 1b.

The histogram of the S-image (see Fig. 2a.) shows the distinct modes corresponding to each of the regions in the blood-smear. The WBC-nucleus can be easily identified by the high values of saturation. In most cases, the WBC-cytoplasm occupies the next level of saturation. However, the ambiguity can be resolved using the spatial information that the cytoplasm is in immediate contact with the nucleus. The image is converted to its HSV equivalent using the following equations:

$$H = \cos^{-1}\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{\frac{1}{2}}}\right] \quad (1)$$

$$S = 1 - \frac{3}{R+G+B}\min(R,G,B) \quad (2)$$

$$V = \frac{1}{3}(R+G+B) \quad (3)$$

Each pixel in the image is represented by a vector of 3 components, namely H, S and V. Since the S-component plays a more conspicuous part, we have weighted it by a factor of 2, while the other two features are given unit weightage. K-Means clustering is performed on this collection of vectors. We have used 6 clusters in our experiments. The centroids are initialized by finding the mean vector and looking for those K-vectors that are farthest from the mean. Euclidean distance in the feature space is used as the measure of dissimilarity. The convergence criteria is that the

difference in the centroids in successive iterations is less than a pre-defined threshold. At the end of this run, we get a class label for each of the pixels, and the centroids for each of the classes.

A priori knowledge helps us conclude that the centroid with maximum saturation corresponds to nucleus. We then crop a rectangular region, surrounding the nucleus, of sufficient area so as to enclose the entire cell. Thus a set of sub-images, each containing only one WBC, is obtained.

Further processing of each of the sub-images involves two steps: (i) Initial estimation of parameters using K-means (ii) Refinement of parameters using EM

2.1 Initial Estimation using K-Means

Each sub-image is separately processed. First, the image is converted to its HSV components. K-means clustering is carried out on the HSV-vectors. Repetition of the clustering step on the small data set results in tighter clusters within the region. We obtain a class label for each of the pixels, and the centroids for each of the classes.

We model each of the clusters by a Gaussian distribution. The initial values of the parameters of the normal distribution can be computed using the clusters obtained by the K-means algorithm. For the kth cluster, the mean is given by:

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i \tag{4}$$

where, x_i is every 3-D vector that belongs to the kth cluster, μ_k is the mean vector and n_k is the number of vectors in the kth cluster.

Since the three features H, S and V are independent, the off-diagonal elements of their covariance matrix can be taken as zero. Hence only the self-covariance of each of the dimensions need to be computed. For the kth cluster, the dth diagonal element of the covariance matrix is given by:

$$C_{dd}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{id} - \mu_{kd})^2 \tag{5}$$

where, n_k is the number of vectors in the kth cluster, x_{id} is the d-th dimension of the ith vector and $\hat{\mu}_{kd}$ is the dth dimension of the mean vector of cluster k.

The values of centroids and variances obtained from the K-means step are used as the initial estimates of the parameters. These values are refined in the subsequent step. The EM algorithm [8] is employed as follows.

2.2 Parameter-refinement using EM

The EM algorithm consists of two major steps: an Expectation step, followed by a Maximization step. The Expectation is with respect to the unknown underlying variables, using the current estimate of the parameters and conditioned upon the observations. The Maximization step then provides a new estimate of the parameters. These two steps are iterated until convergence. The following sub-sections explain in detail the E and M steps in our algorithm.

2.2.1 The E Step:

The E step computes the probability S_{ik} associated with labeling the ith pixel, x_i as belonging to the kth cluster,

$$S_{ik} = \frac{1}{2\pi |\mathbf{C}^k|^{\frac{3}{2}}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\mathbf{C}^k)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)} \tag{6}$$

where, C^k is the covariance matrix associated with cluster k, μ_k is the mean vector of cluster k, i and k take values $1, 2 \ldots N$ and $1, 2 \ldots K$, respectively. Here $N = width \times height$ and K =Number of clusters.

2.2.2 The M step:

The M-step refines the model parameters given the clustering arrived at E-step.

The weighted mean of the kth cluster is updated as:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n_k} S_{ik} x_i}{\sum_{i=1}^{n_k} S_{ik}} \tag{7}$$

The weighted self-correlation of the dth feature in the kth cluster is updated as:

$$\widehat{C}_{dd}^k = \frac{\sum_{i=1}^{n_k} S_{ik} (x_{id} - \hat{\mu}_{kd})^2}{\sum_{i=1}^{n_k} S_{ik}} \tag{8}$$

where, x_{id} is the dth dimension of the ith vector and $\hat{\mu}_{kd}$ is the dth dimension of the mean vector of cluster k.
Both E and M-steps are carried out iteratively. The convergence criteria is taken as,

$$|\hat{\mu}_k^{(n+1)} - \hat{\mu}_k^{(n)}| < Threshold \tag{9}$$

Thresholding each of the distributions results in one region being captured in each distribution. Our a priori knowledge of the relevant regions helps us associate them with the Gaussian distributions obtained. As mentioned earlier, the nucleus-region has the highest values of saturation. Hence the Gaussian distribution whose mean vector has the highest saturation component is identified as corresponding to the nucleus. To find the cytoplasm, we look for the cluster with maximum number of pixels in immediate contact with the nucleus. Fig. 2. compares the histogram of a typical S-component image and the corresponding Gaussian fit. The model parameters are obtained by the EM algorithm.

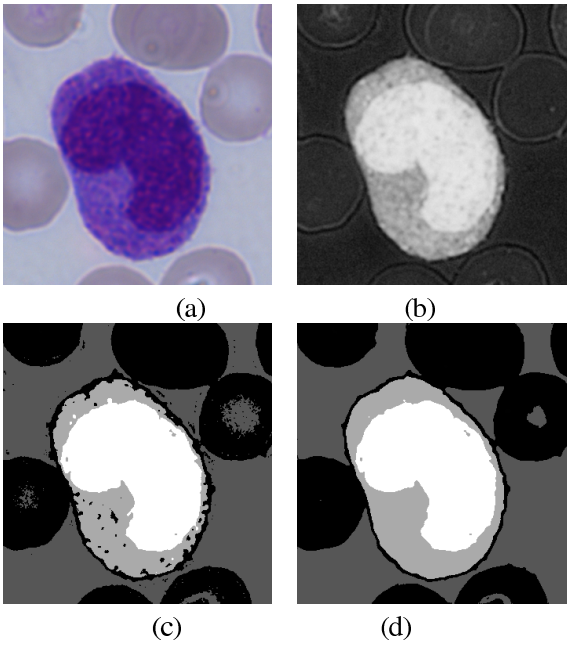


Figure 3: (a) Input Image (b) Saturation Image (c) K-Means Output (d) EM-Output

Figure 4: (a) Protrusions of Neighboring cells (b) Image cleared using Connected component analysis

The entire sequence of processing of a typical image is shown in Fig. 3. The output may need a smoothening process to eliminate specks of mis-classification, if any. This can be accomplished using morphological operations of open-close. Besides, stray instances of platelets might appear, whose coloring pattern resembles those of the WBCs. These are eliminated on the basis of the minimum expected size. The cropped rectangular patch might have protrusions of neighboring cells along with the cell in the center, as illustrated in Fig. 4(a). To eliminate these protrusions, connected component analysis is performed. This helps us retain the cell in the center and ignore the rest (see Fig. 4(b)).

3 Results

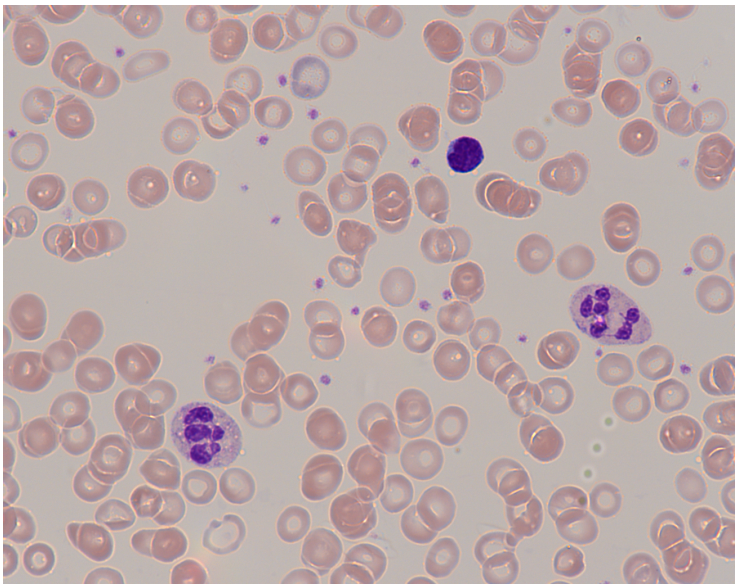
The proposed scheme has been applied on 115 peripheral blood smear slides, stained using May-Grunwald-Giesma (MGG) stain got from the collaborating clinic of the University of Kaiserslautern, Germany. Typical size of the images handled is 1000×1300. Shown in Fig. 5(a) is an image of a blood smear containing 2 neutrophils and a lymphocyte. The segmented outputs obtained are illustrated in Figs. 5, 6 and 7. The ratio of the cytoplasm-pixels to that of the entire image is very low. To avoid the possible merging of the cytoplasm with a more-dominant cluster, we choose the number of clusters beyond the obvious ones, which are the RBCs, background, WBC-cytoplasm and the WBC-nucleus. As can be seen, the image doesn't exhibit very good contrast between the background and the cytoplasm of the WBCs. Our technique successfully segments the image as shown by the outputs.

For cells with granules in the cytoplasm, it is observed that the granules don't get colored homogeneously. The shading causes the lower ends of the granules to be clustered together, and the higher ends of the granules to be separately clustered. However, smoothening of the output helps us recover the entire cytoplasm. We have obtained a segmentation accuracy (manual segmentation carried out by an expert, is taken as the reference) of about 80% on our image dataset of 115 images containing various types of cells, with varying degrees of color contrast between the cells and the background.

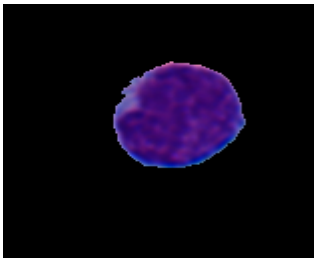
For our calculations, all input values are between 0 and 1. It takes about 20 iterations for the EM-algorithm to converge to an error threshold of 0.00001. In cases where two non-touching cells appear in the same rectangular patch, care is taken to retain only one at a time. However, if the cells happen to be touching, our system doesn't distinguish them as two different cells. This would need the cells to be recognized as clustered, and a declustering technique needs to be subsequently used.

4 Conclusions

We have developed an efficient automatic system for blood cell segmentation from color images of blood smears. This system requires no user-interaction or parameter tuning, which clearly places it above most techniques conventionally used. The system can be easily adapted for any given data set with a known magnification. We utilize the fact that the nucleus exhibits maximum saturation for locating the WBC cells. The system works even when the contrast between the background and the cytoplasm is not perceptible. The performance is good even in cases where the nucleus is multi-lobed, as in neutrophils. However, the technique needs to be enhanced to handle clustered cells, to be clinically used.

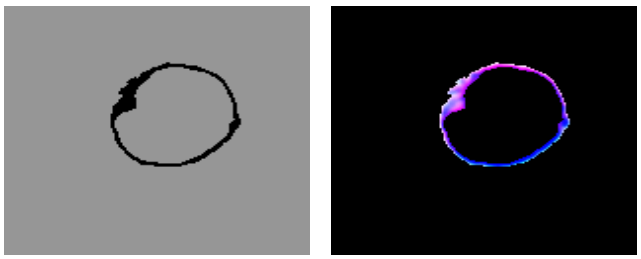


(a)



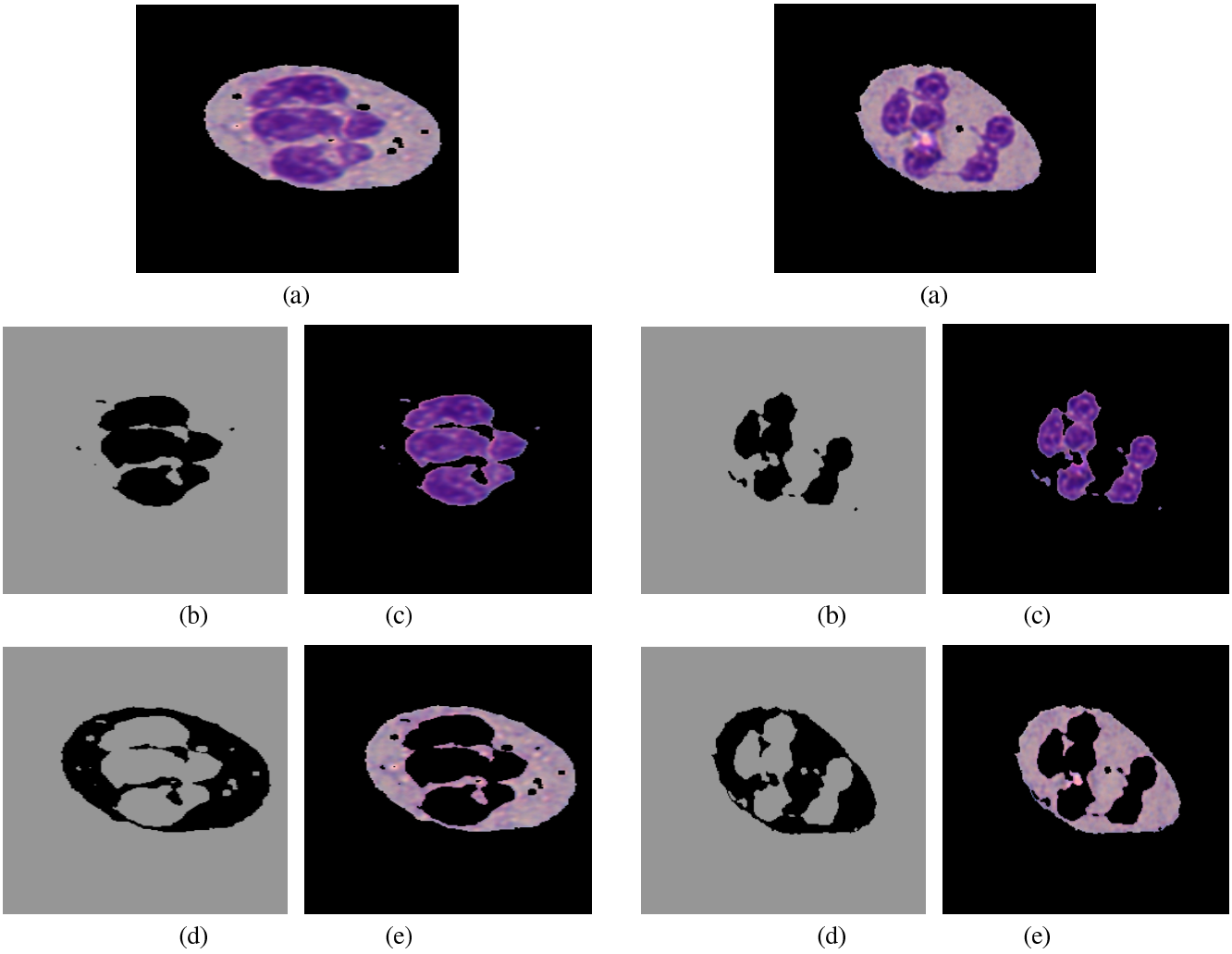
(b)

(c) (d)



(e) (f)

Figure 5: (a) Input Image of the Blood Smear (b) Cropped image of one of the cells (c) Nucleus Mask (d) Cell-Nucleus (e) Cytoplasm Mask (f) Cell-Cytoplasm



(a) (b) (c) (d) (e)

Figure 6: (a) Segmentation Outputs-Example 2

(a) (b) (c) (d) (e)

Figure 7: (a) Segmentation Outputs-Example 3

5 Acknowledgment

The authors thank Prof. M. Pandit and Prof. Link of University of Kaiserslautern, Germany, for providing the data.

References

[1] Cell image segmentation for diagnostic pathology. www.caip.rutgers.edu/riul/research/papers/ps/cell.ps.gz.

[2] Single white blood cell extraction in low resolution. http://sun16.cecs.missouri.edu/jpark.

[3] I. Cseke. A fast segmentation scheme for white blood cell images. In *11th IAPR Int. Conf. on Pattern Recognition, Conf. C: Image, Speech and Signal Analysis*, volume 3, pages 530–533, 1992.

[4] D.Wermser, G.Haussman, and C. Liedtke. Segmentation of blood smears by hierarchical thresholding. In *Computer Vision,Graphics and Image Processing*, volume 25, pages 151–168, 1984.

[5] A. R. Katz. Image analysis and supervised learning in the automated differentiation of white blood cells from microscopic images. Master's thesis, Department of Computer Science, RMIT, Feb. 2000.

[6] B. R. Kumar, D. K. Joseph, and T. V. Sreenivas. Teager energy based blood cell segmentation. In *14th Intl. Conf. on Digital Signal Processing*, pages 925–928, 2002.

[7] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. on System Man and Cybernetics*, 9(1):62–66, Jan. 1979.

[8] Y. Weiss. Motion segmentation using EM- a short tutorial. Technical report, MIT, MA 02139, USA, Nov. 1996. E10-120.

[9] V.A.Kovalev,A.Y.Grigoriev, and H-S. Ahn. Robust Recognition of White Blood Cell Images In *13th Intl. Conf. on Pattern Recognition*, pages 371–375, 1996.