Blood Cell Segmentation and Classification

A Thesis

Submitted for the Degree of

Master of Science (Engineering)

in the Faculty of Engineering

by

Neelam Sinha

Department of Electrical Engineering

Indian Institute of Science

Bangalore – 560 012

August 2002

Synopsis

Medical Image Analysis is the application of Image Processing on medical images, such as X-rays, MRI images, Images of Blood Smears. The main objective is to build a system which, when fed with these images, can identify and understand the patterns for medical diagnostic tasks. Generally human intervention would be needed to interpret the images, arrive at statistics and then make decisions regarding the position of the subject or the course of treatment to be offered. This process would need considerable amount of time, effort and care. It would inevitably be subject to human-errors. The attempt here is to subsitute human effort by a system, which is robust, reliable and automatic. Here, we focus on images of blood smears. Stained blood smears on slides are captured on a camera with specified resolution, resulting in the coloured images. The final goal is to build an automatic system that can diagnose Leukaemic diseases, given these images. The first step towards this would be to work on healthy white blood cells(WBCs) rather than diseased ones. There are nearly 17 classes of WBCs identified through the cycle of Haemopoiesis (formation of blood cells). For now, this project aims at distinguishing between the five major classes of mature cells among WBCs, which include Lymphocyte, Monocyte, Eosinophil, Basophil and Neutrophil. The input to the system is a digital image of the blood smear and it should give the differential count of the types of cells as output.

The main stages of the proposed system are: (i)Acquistion (ii)Segmentation (iii)Feature Extraction (iv)Classification

Acquisition is the process of capturing the blood smear on the slide, into an image, using a digital camera. This decides the quality of the input.

Segmentation is a very crucial step, since the subsequent analysis depends on this. In this stage, the white blood cells are extracted from the complicated background consisting of rbcs, plasma, platelets and cell-fragments. Further, distinction between cell-parts (Nucleus and Cytoplasm) should also be accomplished. A method for segmentation of cells from color images of blood smears in the frame work of statistical modeling is used.

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 80most important feature of this technique is that there are no parameters to be tuned by the user.

Next, features of the cell-parts are computed. Features are based on shape, color, texture and their relative statistics. The features include : Eccentricity, Compactness, Average color of the cytoplasm and nucleus separately, area-ratio between the cytoplasm and the nucleus, and the number of lobes that make the nucleus. The texture features are energy, entropy and correlation, derived from the co-occurence matrix and Coarseness and Busyness, derived from the autocorrelation matrix. A set of library patterns consisting of 50 cells, with about 10 representatives of each of the 5 classes, is created. The limited number of training samples is a big hurdle since not all variations within each class can be covered.

Among the classifiers, Neural Network results in the best accuracy of about 97% on a test population of about 35 samples, with fair representation from all of the classes. Other classifiers such as SVM and Bayes have also resulted in an accuracy of over 80%.

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical Image Analysis aims at building systems that are capable of the requisite handling of medical images. This could involve accomplishment of a wide spectrum of tasks, such as image enhancement, to increase the clarity of the regions of interest as with the low-contrast MRI images (Bhanu's)or texture analysis as in prognosis of ultrasound foetal images for determination of lung maturity (Bhanu's) or recognition of patterns as in Bone Image Analysis needed for forensic applications(Bone Image Segmentation). Detailed studies in human anatomy and complicated surgeries might need simulations requiring visualisation of anatomy.

Computerised eye check-up is a routine application in everyday life. An instance of such a task is in analysis of Blood smears .. The objective of Analysis of Blood smears is to determine the differential count of the white blood cells..(to say,how many of what type) Computerized medical image analysis and visualization algorithms are therefore of fundamental importance to make possibly full use of the information buried in the acquired enormous flood of image data. A group led by Kenan Professor Stephen M. Pizer researches methods related to the analysis of medical images in terms of the objects in them and their shape. Application objectives include the extraction of objects, registration of images based on object matching, measurement of object shape change, and 3D display of objects. Application areas include planning and verification of radiotherapy, planning and delivery of neurosurgery, thoracic surgery, and biopsy, and diagnosis of

schizophrenia via measurement of shape change of brain organs. Angiogram is an X-ray image of the blood vessels network. It might have shadows created due to bodily tissues, such as bones. These shadow regions are obstacles on the angiogram image and therefore a great increase in quality of angiogram can be achieved if those shadows are removed. Better quality of angiogram can help doctors to create better diagnoses and ensure greater probability of correct diagnose.

ultrasonic imaging... Computerised Tomography... Functional MRI...

Medical image analysis tries to find descriptions of important diagnostic signs in medical images and to use them for diagnosis. Computer-derived descriptions are often more accurate and repeatable than those derived by human observers. The descriptions can include shape, colour, pattern, texture, and many other image features. images of skin moles, to extract descriptions which help in diagnosis of melanoma, a skin cancer. images of the interior of the eye

The increasing availability of computing power as well as appropriate modeling and description methods recently enabled rapid development in: physically based simulation of complex medical systems and the underlying biological processes to support therapy planning as well as medical education and training in general; quantitative image analysis for disease and therapy monitoring, including morphological measurement methods for the characterization of biological shape.

For such applications, manual analysis becomes questionable not only because of the amount of work, but also with regard to precision and the poor reproducibility of the results. The motivation to automate stems from the fact that, manual work besides being time-consuming, depends on several conditions, like , the technician's experience, the sample complexity and fatigue. Consistency too can't be guaranteed. Hence the need for automatic medical diagnosis systems is unquestionable.

White cell count (WBC) is the total number of leukocytes in a volume of blood, expressed as thousands/l.

Complete Blood Count:

Differential Blood Count: The WBC count can be done by manual methods or by

automated cell counters. This aims at arriving at the count of each of the cell types.

Manual WBC counting The manual procedure, using a Unopette system and a

Classification and counting of white blood cells (WBC or leukocytes) is the widespread technique of disease diagnosis [1, 2, 3, 4]. Recently, a range of automated WBC counters have cropped up. From point of view of imaging technology, they can be conditionally categorized into three classes: flow cytometry, imaging-based flow microscopy, and pattern recognition based slide readers. The most counters of first class utilize so-called Coulter principle of impedance measurement for a liquid-dispersed blood flow [5, 6, 3]. Some other flow systems use differential enzyme histochemistries, optical measurements of light scatter and cell fluorescence [7, 8, 9]. /**********/ Flow cytometry is a means of measuring certain physical and chemical characteristics of cells or particles as they travel in suspension one by one past a sensing point. In one way, flow cytometers can be considered to be specialised fluorescence microscopes. The modern flow cytometer consists of a light source, collection optics, electronics and a computer to translate signals to data. In most modern cytometers the light source of choice is a laser which emits coherent light at a specified wavelength. Scattered and emitted fluorescent light is collected by two lenses (one set in front of the light source and one set at right angles) and by a series of optics, beam splitters and filters, specific bands of fluorescence can be measured. We can measure physical characteristics such as cell size, shape and internal complexity and, of course, any cell component or function that can be detected by a fluorescent compound can be examined. So the applications of flow cytometry are numerous, and this has led to the widespread use of these instruments in the biological and medical fields.

/**********/ All of these methods dominate on the haematology practice, however, forfeit the rich amount of information available in a visual image.

1.1 Conventional Methods

Blood samples are typically analysed in three or five- part differential cell counters. Based on cell distribution, cell morphology and other indicators, the cell counter can trigger an alarm that the sample is abnormal in some respect and needs manual microscopic analysis. Depending on the degree of automation, a stained and wedged slide can be prepared manually or the sample can be sent forward to a slide maker/stainer that automatically produces a stained slide for the manual differential count. The smear is then inserted individually into the microscope, immersion oil is applied and the medical technician starts looking for an optimal monolayer (examination area) to perform the differential count. The technician will then typically count 100 leucocytes. For each cell the technician finds, the fine focus is adjusted, the cell is studied in detail, the cell type is determined and recorded on an external counting device and then the technician moves on to the next cell until 100 cells have been identified. The time for this procedure varies with the individual technician's skills, the leucocyte concentration and sample type, but is typically in the range of four to 20 minutes. There are several drawbacks to this conventional method.

Q u a l i t y The quality of the results of manual differential counting varies with the number of cells counted, the number of cells skipped, the person's skills and experience, workload and stress level.

1.2 Automation Techniques

Works on Segmentation:

Gray-scale imgs. are less sensitive to variations of lighing conditions and staining quality and need less processing time and storage than do color imags. But we lose out on the advantages that color can offer... It's hence a compromise between complexity, ..

Some of the works only aim at differentiating the Wbcs from the erythrocytes and platelets.They don't look into the cell-structures. $\text{cell}_identify.pdf : Worksongray - scaleimages.The3typesofcells, Wbcs, erythrocytesandplateletsaredifferentiatedbasedonsize, shapean$

Works on Gray-scale:

$cell_classify.pdf : Works on colored images. The 3 types of cells, Wbcs, erythrocytes and platelets are dif$

ieee2.pdf: Only to locate WBCs.Works on colored images.... Thresholding combined with morphological operations for initial labels,based on priori information. The labels are adjusted with a shape detection method based on large regional context information of the cell structure. Circle-identification on the edge-image...to get orbicular shape of the cell.

cseke.pdf: On gray scale images....Healthy cells WBCs are localized based on the apriori information of the blood smears reagrding their approximate size and the range of gray levels. Further processing carried is out on the subimages,each assumed to be containing a single WBC aims at distinguishing the cytoplasm,nucleus and the Background.This is done using Otsu[Ref] thresholding, which maximizes the inter-class variance. The drawbacks are that it would work only if the images have uniform coloring pattern and magnification. Besides, Cytoplasm is not differentiated from the Red Blood Cells Touching cells are not tackled .

$fuzzy_seg.pdf : This uses watershed algorithm that segments based on spatial proximity and intensity hor$

$Fuzzy_detection.pdf : Only to locate WBCs...Gray-scale images..Attempts to take into account the geom$
$level Texture decide...highly subjective labelling...Some more to go....$

Works on Segmentation and Classification: bikhet.pdf: Normal blood cells..Gray-scale images... Shape analysis based on morphological characteristics of their contour and nuclei. Cell Localization by Thresholding.Preprocessing for removal of noise using a median filter. It's further strengthened by enhanced edges. The features used are Area of the Cell,Area of the Nuk,Area of the Cyto, D is the ratio of Nuk area to cell area,Avg color of the cytoplasm, Ratio of Nuk area to Circumference,Cell circularity,Nucleus circularity, zero crossing..?? Granularity and Texture are not exploited. Overlapping and touching cells not resolved.

Turkish.pdf: Works on Coloured

1.3 System Overview

Stained blood smears on slides are captured on a camera with specified resolution, resulting in the coloured images. The final goal is to build an automatic system that can diagnose Leukaemic diseases, given these images. The first step towards this would be to work on healthy white blood cells(WBCs) rather than diseased ones. There are nearly 17 classes of WBCs identified through the cycle of Haemopoiesis (formation of blood cells). For now, this project aims at distinguishing between the five major classes of mature cells among WBCs, which include Lymphocyte, Monocyte, Eosinophil, Basophil and Neutrophil. The input to the system is a digital image of the blood smear and it should give the differential count of the types of cells as output.

The main stages of the proposed system are: (i)Acquistion (ii)Segmentation (iii)Feature Extraction (iv)Classification

Acquisition is the process of capturing the blood smear on the slide, into an image, using a digital camera.

Segmentation is a very crucial step, since the subsequent analysis depends on this. In this stage, the white blood cells are extracted from the complicated background consisting of rbcs, plasma, platelets and cell-fragments. Further, distinction between cell-parts(Nucleus and Cytoplasm) should also be accomplished.

Next, features of the cell-parts are computed. Features are based on shape, color, texture and their relative statistics. The features used are: Eccentricity, Compactness, Average color of the cytoplasm and nucleus separately, area-ratio between the cytoplasm and the nucleus, and the number of lobes that make the nucleus. The feature-set must A set of library patterns consisting of 50 cells, with about 10 representatives of each of the 5 classes, is created. The limited number of training samples is a big hurdle since not all variations within each class can be covered.

Among the classifiers, Neural Network

1.4 Characteristics of Cells

White blood cells (leucocytes; Gr. leukos: white) are the immune and defence cells of the blood. They can be divided into several classes:

granulocytes lymphocytes monocytes

Granulocytes : are named for their prominent and characteristic cytoplasmic granules as seen in standard blood smears (Romanovsky stain) are also known as polymorphonuclear leucocytes (polymorphs) because their nuclei have several lobes can be divided into neutrophils, eosinophils and basophils.

Neutrophils make up approximately 60have segmented nuclei with 2-5 lobes have mostly pale granules and some darker-staining lysosomal granules.

Eosinophils

make up 1-3have bilobed nuclei have bright red lysosomal granules

Basophils make up fewer than 1have bilobed nuclei often obscured by granules have deep blue granules

Lymphocytes make up 20-30mostly have round, condensed nuclei typical of cells with little biosynthetic activity; cytoplasm forms only a narrow rim around the nucleus. Monocytes

make up 4-10are large cells with indented nuclei

1.5 Organisation of Thesis

1.6 Conclusion

Chapter 2

Blood Cell Segmentation using EM algorithm

Summary:

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.

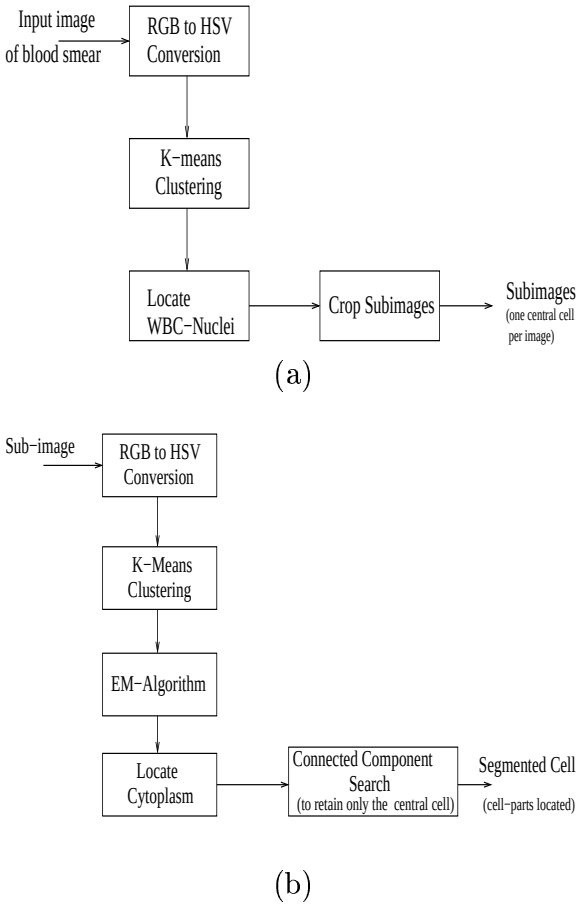


Figure 2.1: System Overview (a) Stage 1 (b) Stage 2

2.1 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.* (dor,) 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 (par,) 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.* (Kumar et al., 2002). 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 (Katz, 2000) 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.* (D.Wermser et al., 1984) have introduced a hierarchical thresholding scheme using a priori information regarding chromatic properties of background and cell components. Kovalev *et al.* have proposed a three- step algorithm to segment white blood cells, employing prior knowledge of color information and using a circle-shaped approximation, Cseke (Cseke, 1992) investigated the multi-step segmentation scheme, which implements the automatic thresholding method suggested by Otsu (Otsu, 1979).

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.

2.2 Segmentation scheme

Figure 2.1 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:

$$H = \cos^{-1}\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{\frac{1}{2}}}\right] \tag{2.1}$$

$$S = 1 - \frac{3}{R+G+B}\min(R,G,B) \tag{2.2}$$

$$V = \frac{1}{3}(R+G+B) \tag{2.3}$$

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

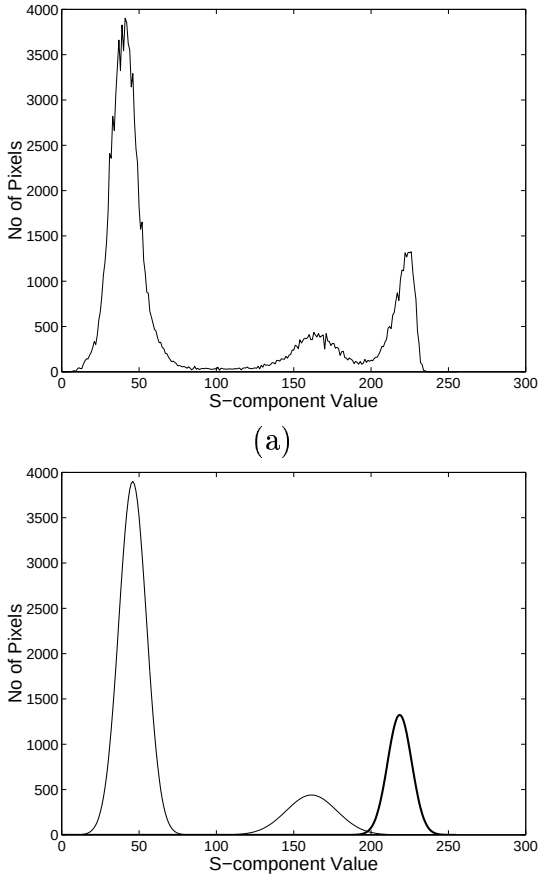


Figure 2.2: (a) Original histogram (b) corresponding Gaussian fit

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: (1)Initial estimation of parameters using K-means (2)Refinement of parameters using EM

2.3 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:

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i \tag{2.4}$$

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:

$$C_{dd}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{id} - \mu_{kd})^2 \tag{2.5}$$

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 (Weiss, 1996) is employed as follows.

2.4 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.

2.4.1 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,

$$S_{ik} = \frac{1}{2\pi|\mathbf{C}^k|^{\frac{3}{2}}} e^{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\left(\mathbf{C}^k\right)^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)} \tag{2.6}$$

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.

2.4.2 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:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n_k} S_{ik} x_i}{\sum_{i=1}^{n_k} S_{ik}} \tag{2.7}$$

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

$$\widehat{C}_{dd}^k = \frac{\sum_{i=1}^{n_k} S_{ik}(x_{id} - \hat{\mu}_{kd})^2}{\sum_{i=1}^{n_k} S_{ik}} \tag{2.8}$$

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,

$$|\hat{\mu}_k^{(n+1)} - \hat{\mu}_k^{(n)}| < Threshold \tag{2.9}$$

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. Figure 2 compares the histogram of a typical S-component image and the corresponding Gaussian fit. The model parameters are obtained by the EM algorithm.

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)).

2.5 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.

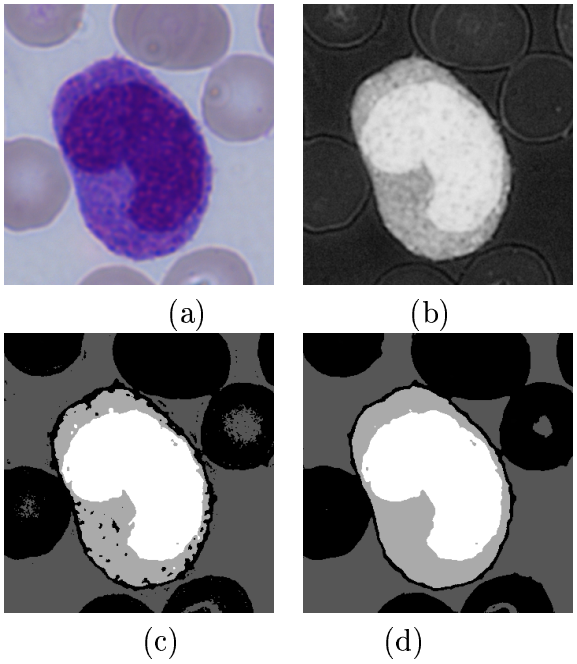


Figure 2.3: (a) Input Image (b) Saturation Image (c) K-Means Output (d) EM-Output

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

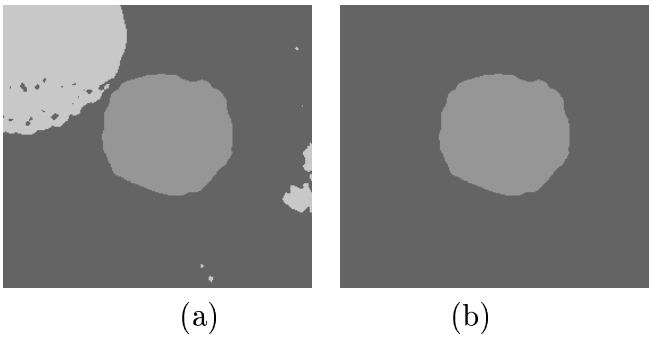


Figure 2.4: (a) Protrusions of Neighboring cells (b) Image cleared using Connected component analysis

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 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.

2.6 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. 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.

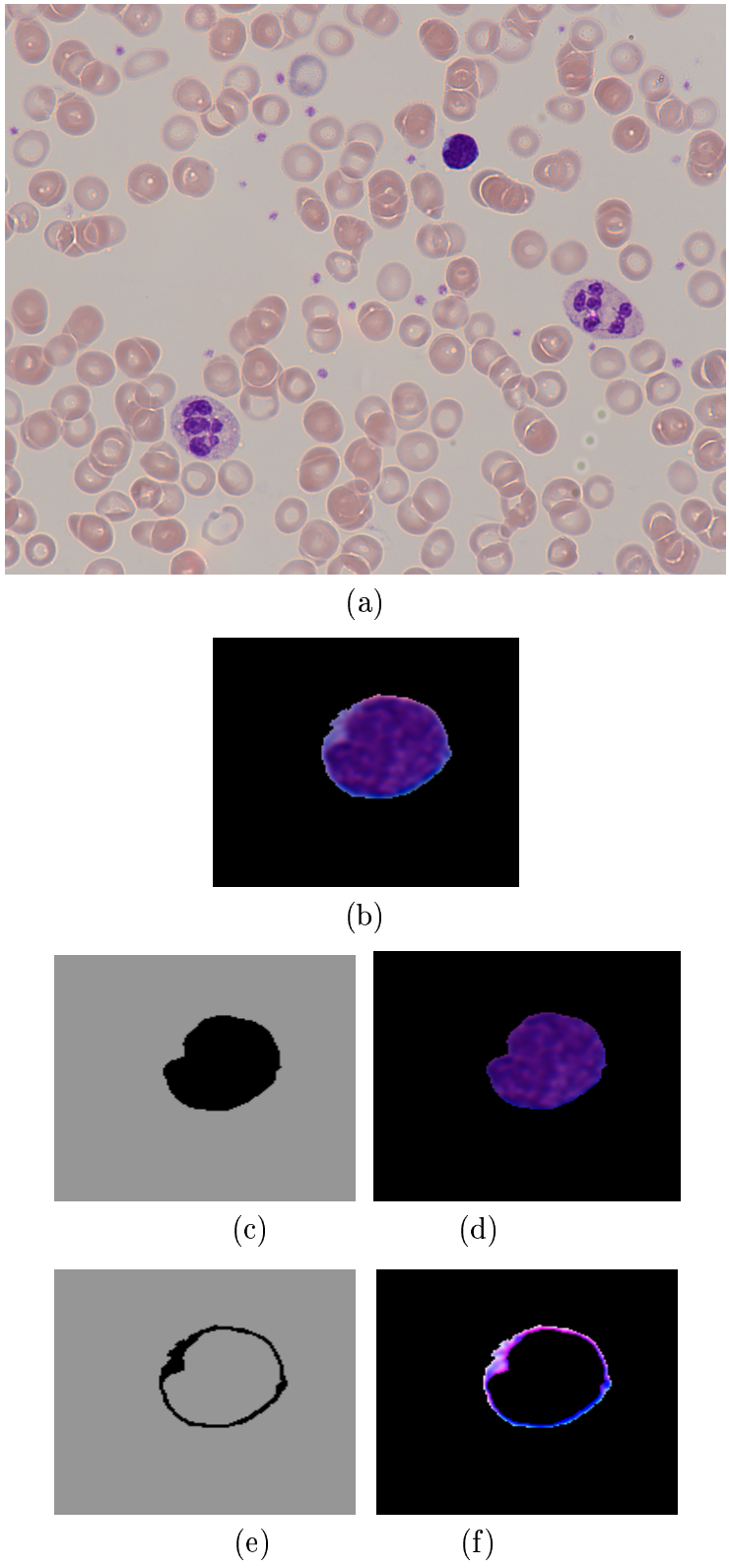
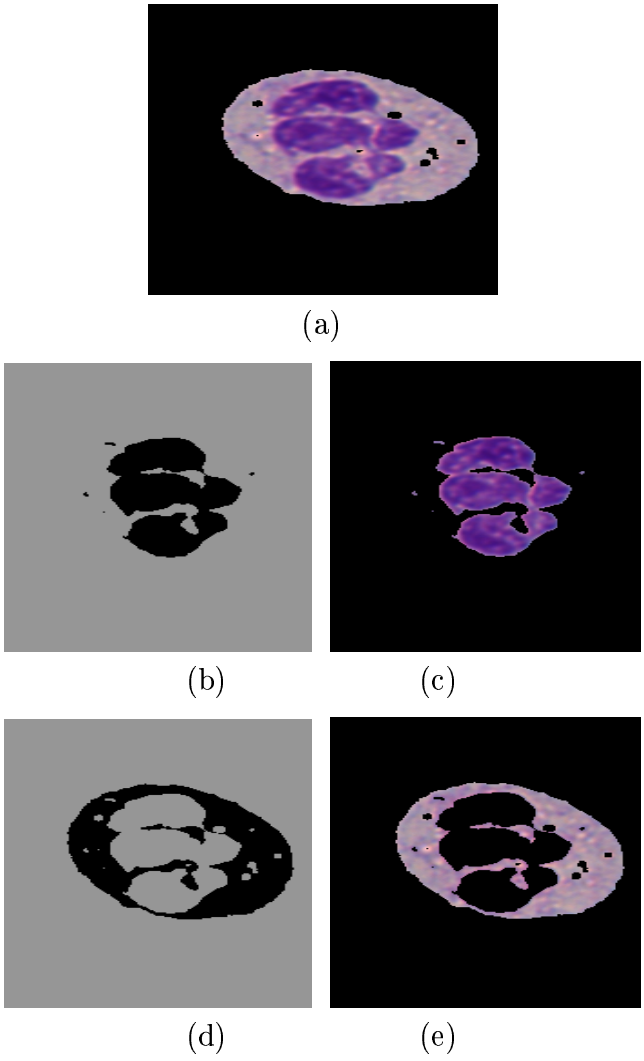
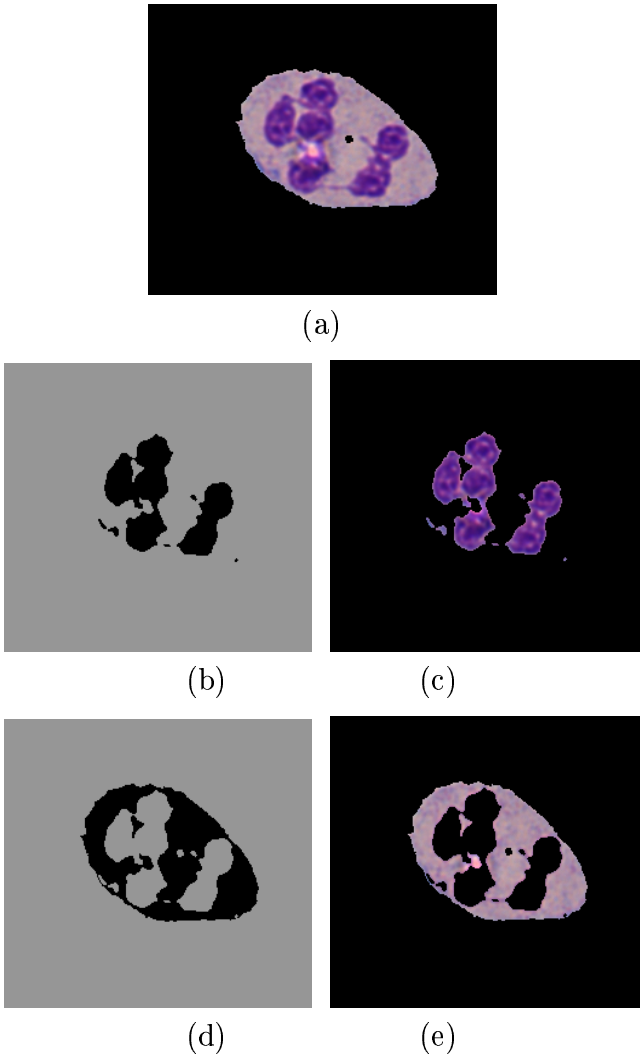


Figure 2.5: (a) Input Image of the Blood Smear (b) Cropped image of one of the cells (c) Nucleus Mask (d) Cell-Nucleus (e) Cytoplasm Mask (f) Cell-Cytoplasm



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

Figure 2.6: (a) Segmentation Outputs-Example 2



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

Figure 2.7: (a) Segmentation Outputs-Example 3

Chapter 3

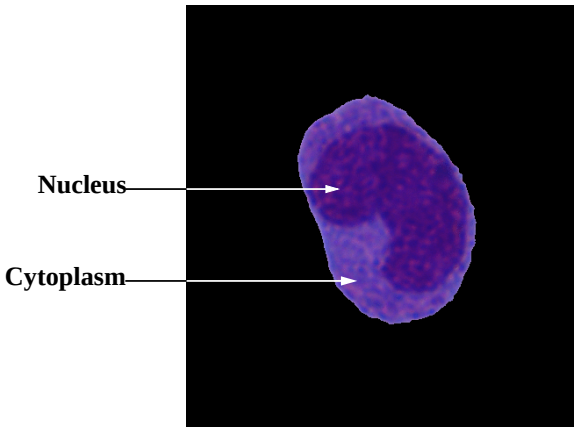
Feature Extraction

3.1 Introduction

Features are representative measures of a pattern and are chosen by their ability to reconstruct the input pattern. According to Kittler and Devijver,the process of feature extraction is defined as, "extracting from raw data the information which is more relevant for classification purposes, in the sense of minimizing within class pattern variability while enhancing between class pattern variablity". It is seen as a process of mapping original features into features that are more useful. It is construction of a vector in a multidimensional space, given a data set (such as an image), where each dimension represents an attribute of the image that is believed to carry information useful in the classification of that image. The optimal set of features must capture the distinguishing characteristics of each class and maintain good class-separability.

The design of the feature extractor is very problem dependent. The ideal feature extractor would produce the same feature vector for all patterns in the same class, and different feature vectors for patterns in different classes. In practice, different inputs to the feature extractor will always produce different feature vectors. But we look for those features that result in small within-class variability relative to the between-class variability.

The advantages in working with features rather than the images as a whole:



Cell-parts

- Computational complexity of pattern classification is reduced by handling the data in the lower dimensional feature space resulting in more effecient and faster classification .
- Storage space needed for the feature data is far less as compared to that needed by the entire image.

Suitability of a feature is domain dependent, and hence a task specific approach must be followed. Here,

- The features have to be rotation and translation invariant.
- The features must incorporate enough tolerance, acknowledging the extent of variations that could occur across various cross-sections of human population.

3.2 Cell Structure

White Blood Cells comprise of two parts, the central one called the Nucleus, and the one surrounding it called the Cytoplasm,as shown in the fig. The variations of features of these cell parts enable us to distingush between the various classes of White bloood cells. The characteristics of each of the cell-parts separately, or a relative measure between the two, or a combination of both, could be used to classify the cell.

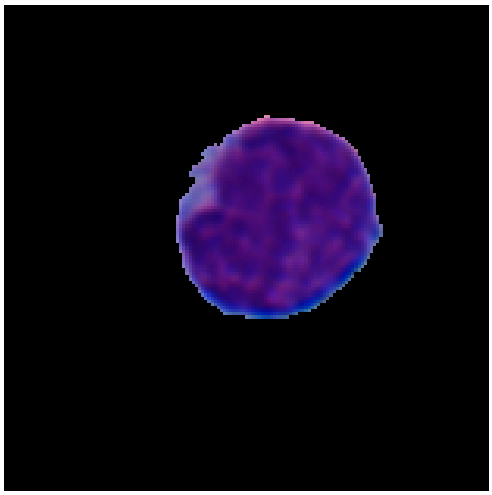
Based on the cytoplasm, cells are broadly classified as Granulocytes and Non-Granulocytes.

Granulocytes exhibit the presence of granules in the cytoplasm, while non-granulocytes don't. The differences in the color, size and spread of the granules serve as vital distinguishing cues among the different types of granulocytes. Granulocytes have a single-lobed nucleus while the Non-Granulocytes generally have multi-lobed nucleus.

The cell-parts are characterized in terms of their shape, color ,texture and relative measures between them.

The Non-Granulocytes are of 2 types.

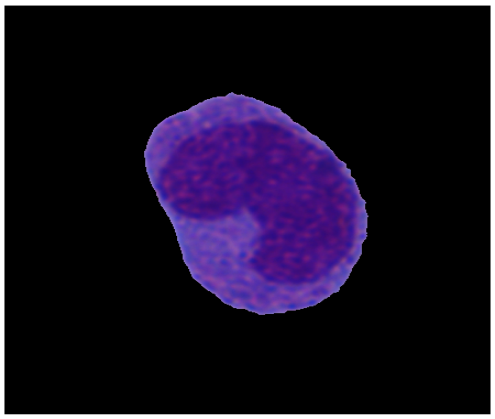
- Lymphocyte: Lymphocytes can be easily identified by the low value of the area-ratio between the cytoplasm and the nucleus, since the cytoplasm is present only along the a thin rim. The circular shape of the nucleus is fairly consistent. The nucleus is single-lobed.



Lymphocyte

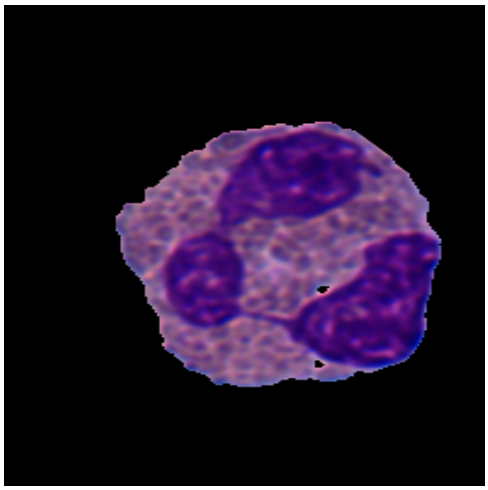
- Monocyte: Monocytes exhibit a dent in the otherwise fairly ellipsoidal nucleus. The shape and size of the dent are not consistent at all. Here too, the nucleus is single-lobed. Unlike Lymphocytes, here the cytoplasm occupies a fair share of the total cell area.

The Granulocytes are divided into 3 classes :



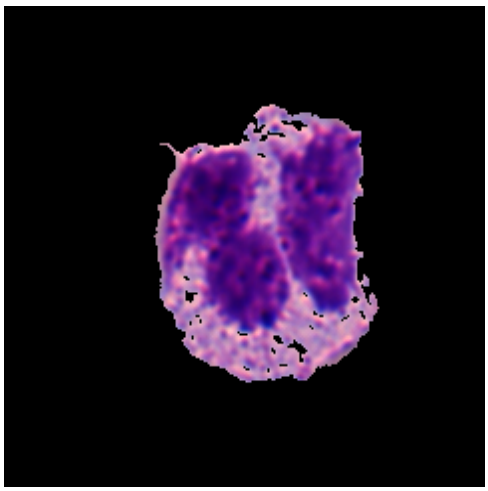
Monocyte

- Eosinophils: Eosinophils have compactly packed small red-colored granules in the cytoplasm. The Nucleus is generally bi-lobed.The shape of the cell boundary is generally oval.



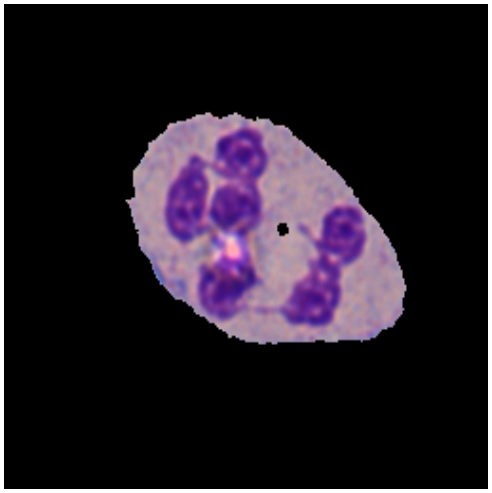
Eosinophil

- Basophils: Basophils have large blue-colored granules loosely scattered in the cytoplasm. They have bilobed nuclei often obscured by granules.The shape of the cell boundary is generally oval.



Basophil

- Neutrophils: Neutrophils have small purple-colored granules that are loosely scattered in the cytoplasm.The Nucleus is generally muti-lobed. The shape of the cell boundary is generally oval.



Neutrophil

3.3 Features used :

We look at features that virtually translate visual cues to numbers.

3.4 Shape features

Shape Descriptors are a set of numbers that are produced to describe a given shape.The shape may not be entirely reconstructable from the descriptors,but they must be distinct enough for different shapes to enable discrimination. To compute these features, the inputs are binary masks of the nucleus and the cytoplasm.

The features used here are:

- Eccentricity

  Eccentricity is defined as the ratio between the major and minor axes. It gives an idea of how close the shape is to a circle. Since the input image is binary,it could be computed as the ratio of the eigen values of the covariance matrix of the position vectors of the foreground pixels.

  $$\widehat{\mathbf{P}}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

  represent the position vectors

  Then their covariance matrix is given by

  $$\mathbf{C} = E\left[\left(\widehat{\mathbf{P}}_i - \mathbf{m}_p\right)\left(\widehat{\mathbf{P}}_i - \mathbf{m}_p\right)^T\right] \tag{3.2}$$

  The eigen values are given by

  $$\mathbf{C}\widehat{\boldsymbol{\nu}} = \lambda\widehat{\boldsymbol{\nu}} \tag{3.3}$$

  for non-zero values of $\widehat{\boldsymbol{\nu}}$

  $$\text{If } \lambda_1 > \lambda_2 \tag{3.4}$$
  $$\text{then, } Ecc = 1 - \frac{\lambda_2}{\lambda_1}$$

- Compactness Compactness is computed as the ratio of area to square of the perimeter. It tells about the extent of indentation of the boundary. The value is high if the boundary is smooth, and low if it is serrated.

$$\text{Compactness} = \frac{\text{Area}}{\text{Perimeter}^2} \tag{3.5}$$

3.5 Relative statistics

Area-ratio is a relative-statistics measure. Area is taken as the count of the foreground pixels.Thus it is the ratio of the number of pixels that make up the cytoplasm to the ones that make up the nucleus.

$$\text{Area-Ratio} = \frac{\text{Pixel count of Cytoplasm}}{\text{Pixel count of Nucleus}} \tag{3.6}$$

3.6 Number of Lobes

To find the number of lobes, a declustering technique involving computation of the negative distance transform, followed by watershed algorithm is used. Distance transform is defined for binary images.It is computed for every foreground pixel, as it's distance from the nearest background pixel. The result of the transform is a greylevel image that looks similar to the input image except that the greylevel intensity of points inside foreground regions are changed to show the distance to the closest boundary from each point. The distance metric chosen here is, Euclidean.However other metrics like Chessboard or Cityblock could also have been used. Distance transform is defined as:

$$D(f) = \{p : p = \min d(f, b)\} \tag{3.7}$$

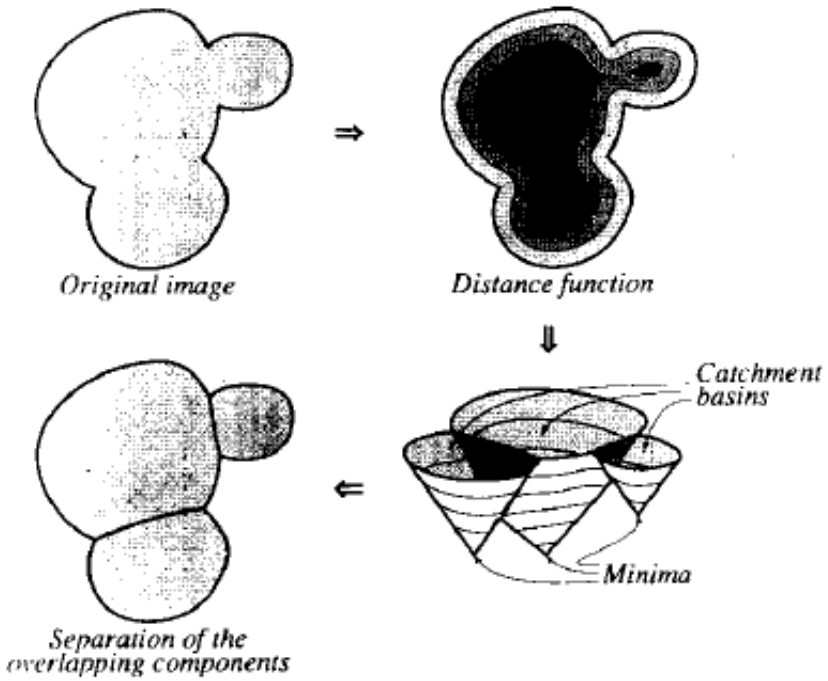
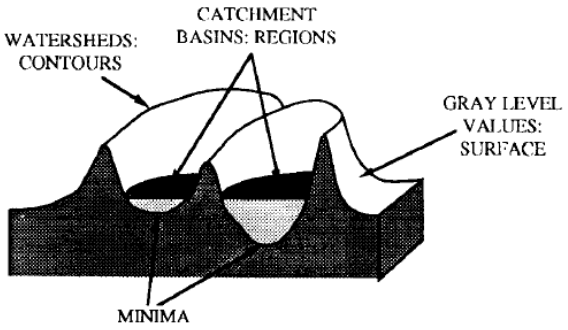


Illustration of the Declustering Technique

for all b that belongs to the image. where, f is the foreground pixel under consideration and b is a background pixel in the image. The negative of this matrix is the Negative distance transform.

$$N(f) = 255 - D(f) \tag{3.8}$$

The watershed transform is a fundamental image segmentation tool in mathematical morphology. Segmentation involves dividing an image up into different regions, with each region having some characteristics in common. Another way of saying this is that, segmentation involves partitioning the image plane so that each partition is homogenous with respect to some criteria. Here,we are looking at gray-level as the criteria. The watershed transformation is based on an analogy with topographic reliefs. An image can be thought of as a three dimensional relief with the greyscale value at each point corresponding to height. Imagine that the relief has holes punched at each regional



Watershed Algorithm

minimum at that it is allowed to sink slowly into some fluid.

The fluid will flow up through the holes and across the relief according to its topography. At some places, two flows of fluid will meet: we erect a barrier at such a point. Once the relief has become completely covered by water, we end up with a structure with several barriers or dams on it. These dams represent so-called watershed lines and serve to mark the divisions between the "catchment basins" of the relief.

One of the main advantages of the watershed transform as a segmentation tool is that the segment boundaries it produces are closed.

The set of the Catchment basins of the greyscale image I is equal to the set $X_{h_{max}}$ obtained after the following recursion:

(a)

$$X_{h_{min}} = T_{h_{min}}(I) \tag{3.9}$$

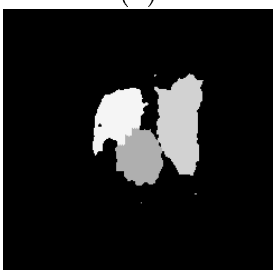
where, $X_{h_{min}} = \{p \in D_I, I(p) \leq h_{min}\}$ where, D_I is the set of values taken by the image I.

b)

$$\forall h \in [h_{min}, h_{max-1}], \qquad X_{h+1} = \min_{(h+1)} \bigcup IZ_{T_{h+1}}(I)\{X_n\} \tag{3.10}$$

(a)

(b)



(c)

where, $IZ_A(B) = \bigcup_{i\in[1:k]} iZ_A(B_i)$ given,

$$iZ_A(B_i) = \{p \in A, \forall j \in [1,k] - \{i\}, \quad d_A(p,B_i) < d_A(p,B_j)\} \tag{3.11}$$

The lobe-centres will appear as the darker regions in the negative distance transform image.The regions that form the linkage between the lobes will appear brighter. The darker regions are set aside as the markers.The markers are chosen well within a predefined size threshold and range of gray levels. These markers are then used for region-growing using watershed. The boundaries obtained by this method make use of the contours of the individual lobes.The quality of declustering doesn't depend on the relative size of the touching lobes. The advantage is that the contour information is not lost, as would have been if one used morphological operations of open-close, for declustering.

3.7 Color features

Here the segmented nucleus and cytoplasm are taken as inputs. The average value of each of the color components(R, G and B components) of the nucleus and that of the cytoplasm are computed.

$$Mean_C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{3.12}$$

where,N is the total number of pixels in the region of interest(either Nucleus or Cytoplasm)

C_i is the corresponding color(either R or G or B) component of the ith pixel

3.8 Texture features

The texture of the cytoplasm across the various classes, is visually distinguishable, but it is not so with the Nucleus. The difficulty in computing texture features is in obtaining a sizeable rectangular patch that can capture the texture. The texture of the nucleus across the various classes cannot be distinguished by the human eye. Texture is defined as a function of the spatial variation in pixel intensities. Image texture has a number of perceived qualities which play an important role in describing texture. Laws identified the following properties as playing an important role in describing texture: uniformity, density, coarseness, roughness, regularity, linearity, direc- tionality, direction, frequency, and phase. The fact that the perception of texture has so many different dimensions is an important reason why there is no single method of texture representation which is adequate for a variety of textures.

Texture is quantified using the following methods:

3.8.1 Statistical Methods

One of the defining qualities of texture is the spatial distribution of gray values.

- Gray-Level Co-occurence Matrix(GLCM) Spatial gray level co-occurrence estimates image properties related to second-order statistics. Haralick [10] suggested the use of gray level co-occurrence matrices (GLCM) which have become one of the most well-known and widely used texture features. The G G gray level co-occurrence matrix P_d for a displacement vector $d = (dx,dy)$ is defined as follows.

$$\mathbf{P}_d = |((r,s),(t,v) : I(r,s) = i, I(t,v) = j| \tag{3.13}$$

The entry (i,j) of P_d is the number of occurrences of the pair of gray levels i and j which are a distance d apart. where $(r,s),(t,v)$.....

texture-review.pdf The co-occurrence matrix reveals certain properties about the spatial distribution of the gray levels in the texture image. For example, if most of the entries in the co-occurrence matrix are concentrated along the diagonals, then the texture is coarse with respect to the displacement vector d. Haralick has proposed a number of useful texture features that can be computed from the co-occurrence matrix. The co-occurrence matrix features suffer from a number of difficulties. There is no well established method of selecting the displacement vector d and computing co-occurrence matrices for different values of d is not feasible.

The features:

1. Energy= $\sum_i \sum_j \mathbf{P}_d{}^2(i,j)$

2. Entropy= $-\sum_i \sum_j \mathbf{P}_d(i,j) \log \mathbf{P}_d(i,j)$

3. Contrast= $\sum_i \sum_j (i-j)^2 \mathbf{P}_d(i,j)$

4. Homogeneity= $\sum_i \sum_j \frac{\mathbf{P}_d(i,j)}{1+|i-j|}$

5. Correlation= $\sum_i \sum_j \frac{(i-\mu)(j-\mu)\mathbf{P}_d(i,j)}{\sigma_x \sigma_y}$

   where μ is the mean of $\mathbf{P}_d$ and σ_x and σ_y are the standard deviations of $\mathbf{P}_d(x)$ and $\mathbf{P}_d(y)$ respectively.

- AutoCorrelation

An important property of many textures is the repetitive nature of the placement of texture elements in the image. The autocorrelation function of an image can be used to assess the amount of regularity as well as the fineness/coarseness of the texture present in the image. Formally, the autocorrelation function of an image I(x, y) is defined as follows:

$$\rho(x,y) = \frac{\sum_i \sum_j I(i,j)I(i+x, j+y))}{\sum_i \sum_j I^2(i,j)} \tag{3.14}$$

This function is related to the size of the texture primitive (i.e., the fineness of the texture). If the texture is coarse, then the autocorrelation function will drop off slowly; otherwise, it will drop off very rapidly. For regular textures, the autocorrelation function will exhibit peaks and valleys.

The features extracted are i) Coarseness ii) Contrast and iii) Busyness.

The features :

1. Coarseness

$$C_s = \frac{2}{\frac{\sum_i \sum_j Max(i,j)}{n} + \frac{\sum_i \sum_j Max(i,j)}{m}} \tag{3.15}$$

   where $Max(i,j) = 1$ if point (i,j) is a either row maxima or column maxima, else $Max(i,j) = 0$.

2. Contrast

$$C_t = \frac{M_a \times N_t \times C_s^{\frac{1}{\alpha}}}{n \times m} \tag{3.16}$$

where M_a is the average module of the gradient of the autocorrelation function. N_t - number of points having module greater than a threshold t. The threshold t is set at the middle point of the range of the module matrix and the value of α is set at 4.

3. Busyness

$$B_s = 1 - C_s^{\frac{1}{\alpha}} \tag{3.17}$$

where $\frac{1}{\alpha}$ is a power to make C_s significant against 1.

3.8.2 Frequency Based Methods

Filter Design : The two dimensional Gabor function is given by

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right) \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi jWx\right] \tag{3.18}$$

$$G(u,v) = \exp\left\{-\frac{1}{2}\left[\frac{(u-W)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right]\right\} \tag{3.19}$$

where $\sigma_u = 1/2\pi\sigma_x$ and $\sigma_v = 1/2\pi\sigma_y$

here $g(x,y)$ is the mother Gabor wavelet and its self-similar filter dictionary is obtained by the appropriate dilations and rotations of $g(x,y)$ through the following generating function:

$$g_{mn}(x,y) = a^{-m}G(x',y'), a > 1, m, n = \text{integer} \tag{3.20}$$

$x' = a^{-m}(x\cos\theta + y\sin\theta); y' = a^{-m}(-x\sin\theta + y\cos\theta)$ where $\theta = n\pi/K$ and K is the total number of orientations. The scale factor a^{-m} in (3.20) is meant to ensure that the energy is independent of m.

In our filter design we have considered four scales and six orientations. The filter parameters σ_u and σ_v are computed from the lower(U_l) and upper (U_h) center frequencies of interest.

$a = (U_h/U_l)^{\frac{1}{s-1}}$, $\sigma_u = \frac{(a-1)U_h}{(a+1)\sqrt{2\ln 2}}$

$\sigma_v = \tan\left(\frac{\pi}{2k}\right)\left[U_h - 2\ln\left(\frac{\sigma_u^2}{U_h}\right)\right]\left[2\ln 2 - \frac{(2\ln 2)^2\sigma_u^2}{U_h^2}\right]^{-0.5}$

the default values of U_l and U_h are set at 0.05 and 0.4 radians respectively.

Feature Representation

The Gabor wavelet transform of a given image $I(x, y)$ is defined as :

$$W_{m,n}(x, y) = \int I(x_1, y_1) {g_{mn}}^*(x - x_1, y - y_1) dx_1 dy_1 \tag{3.21}$$

where $*$ indicates the complex conjugate. In implementation the 2-D convolution is avoided by taking inner product of fourier transform of image $I(x, y)$ and each of the Gabor-wavelet filters $g_{mn}(x, y)$.

The features (mean μ_{mn} and standard deviation σ_{mn}) are obtained by

$\mu_{mn} = \sum_x \sum_y |W_{mn}(xy)|$ and

$\sigma_{mn} = \sqrt{\sum_x \sum_y \left(|W_{mn}(x, y)| - \mu_{mn}\right)^2}$

The resulting feature vector is $\mathbf{f} = [\mu_{00}\ \mu_{01} \ldots \sigma_{00}\ \sigma_{01} \ldots]$

3.9 Conclusion

In our trials, a 16-dimensional feature vector consisting of

- Area ratio

- Number of lobes in the Nucleus
- Eccentricity of Nucleus
- Eccentricity of Cytoplasm
- Compactness of Nucleus
- Mean Red Value of Nucleus
- Mean Green Value of Nucleus
- Mean Blue Value of Nucleus
- Mean Red Value of Cytoplasm
- Mean Green Value of Cytoplasm
- Mean Blue Value of Cytoplasm
- Energy
- Entropy
- Correlation
- Coarseness
- Busyness

The cytoplasm mask has two boundaries, the inner one which coincides with the boundary of the nucleus and the outer one which forms the cell boundary. The inner boundary information is accounted in computations of the nucleus descriptors. As observed in all the classes, the cell boundary is nearly oval. Hence compactness of Cytoplasm is omitted. The Gabor features have been omitted, since enough texture information could not be captured in the maximum available window of size 32 X 32. It resulted in nearly identical features for visibly different textures, belonging to different classes.

Chapter 4

Classification and Results

Classification is the task of assigning to the unknown test vector, a label from one of the known classes.The classifier evaluates the features of the test pattern and decides the label of the test pattern. The features may be individually bad, but their combination must enable correct classification. The optimal feature set should not contain any of those features that confuse the classifier. Further, the redundant features must be eliminated for improved effeciency.

According to the classical rule of thumb, the number of training patterns for each class must be 5 to 10 times the dimensionality of the feature vector. But, due to inavailability of sufficient data we have had to make do with a far smaller training set. The characteristics of the features extracted decide the performance of the classifers. For features with good discriminative power, a linear classifier like Nearest Neighbour can be used. But if the patterns are very close in the feature space, regression methods like Neural network based classifiers must be used. For optimal performance of the system, the features extracted must have the following properties:

- Invariance to tranformations like scaling, rotation and translation, so that all variations of a given pattern are correctly recognised.
- Robustness to noise
- Good discriminative power, ie. patterns from different classes must be well separated

in feature space.

- As low a dimension as possible without sacrificing their discriminative power.

4.1 Supervised learning:

The aim in supervised classification is to construct a model for predicting the class to which an object belongs on the basis of a vector of measurements of that object.Here, "learning" is the ability to improve its performance at classifying the instances presented to it. A good supervised learning algorithm is one that produces hypotheses (classifiers) that do a good job of predicting the class of unseen examples. The quality of classifier decisions is closely related to the quality and amount of information that is available.Hence, the patterns should represent as complex a description as possible.On the other hand, a large number of description features would result.Therefore, the object description is always a trade-off between the permissible classification error, the complexity of classifier construction and the time required for classification.

Curse of Dimensionality: katz-mbc.pdf Whenever a large number of features exist and thus a large number of hypotheses are possible, there is a danger of using the resulting freedom to learn classifiers based on meaningless or irregular attributes of the data set. This problem is a common one in all kinds of learning algorithms and is known as overfitting or overtraining.

4.2 Non-parametric classifiers:

The simplest approach to classifier design is based on a measure of similarity.The patterns exhibiting more similarity among themselves are grouped together. The classification process evaluates the measure of similarity to decide which class a given pattern belongs to. They do not make any assumptions about the data distributions in feature space. To improve the performance of such a classifier the following must be taken care of :

- An appropriate measure of similarity

- The Training set must contain all possible variations within each class.

The most common measure of similarity is the Euclidean distance metric. The simplest of the classifiers belonging to this group is the Nearest Neighbour Classifier.Variations of this classifier like k-NN,... are also used..

Each training instance is regarded as a point in n-dimensional space, where n is the number of features in the feature vector . When a test instance is presented, the euclidean distance from the point represented by the test instance to each training instance is calculated.

4.2.1 Nearest Neighbour

In the simplest single nearest neighbor case (1-NN) the classification of the nearest neighbor (shortest euclidean distance) becomes the classification of the test instance.

4.2.2 K-Nearest Neighbour

When more number of the nearest neighbors (k-NN) are taken into consideration, each of the k nearest neighbors is determined as before. Each of these nearest neighbors is given a "vote" for its classification and the classification with the highest number of votes is assigned to the test instance [Van1, pp175-177].

4.3 Parametric classifiers:

This group of classifiers relies on the probability of class-occurence and the conditional densities corresponding to the pattern and the classes. Conditional probabilites are assumed to take some standard statistical distribution.Once the form of the distribution is assumed, the parameters that define the model best are determined. This is accomplished using the standard techniques like Bayesian learning rule.

4.3.1 Bayesian Classifier

Bayes Classifier(with Gaussian distributions) A parametric classifier that assumes multivariate Gaussian ("Normal") data distributions in the multi-dimensional feature space. The class means and covariance matrix are estimated from the training data, and new samples are labeled as the class having the maximum aposteriori probability. Using a priori information,the aposteriori probability of the occurrence of the class, given a feature vector, is computed. The parameters describing the distribution are not usually known, and leaning must find the estimate of these parameters.

4.4 Classifiers based on geometric appraoches:

These are based on the estimation of the decision boundaries in the feature space, making use of appropriate error-minimizing criteria. Multi-layer perceptron and SVMs

The advantages of these classifiers include:

- Automatic extraction of parameters
- Inherent non-linearity offering approximation to a wide range of decision boundaries
- Trainability

The major disadvantage of these techniques is that an addition of a new prototype requires the repetition of the time-consuming training process.

4.4.1 SVM

It is basically a two class classifier that optimizes the margin between them. Here we find a hyperplane optimally dividing two classes which does not depend on a probability estimation. This optimal hyperplane is a linear decision boundary which separates the two classes and leaves the largest margin between the vectors of the two classes . It is observed that the optimal hyperplane is determined by only a small fraction of the data points, the so-called support vectors. The support vector classifier training algorithm is a procedure to find these vectors.

Bayesian Classifer Results

Class	A	B	C	D	E	Errors
A	6	1	0	0	0	1
B	0	5	1	1	0	2
C	0	0	0	8	0	0
D	0	0	2	0	2	2
E	0	0	2	0	2	2

Table 4.1: Confusion Matrix Color Shape features

Class	A	B	C	D	E	Errors
A	2	1	1	3	0	5
B	0	7	0	0	0	0
C	2	0	2	4	0	6
D	0	0	0	3	1	1
E	0	1	3	3	1	7

Table 4.2: Confusion Matrix Texture features

4.4.2 Neural Nets

Networks of non-linear computing elements (neurons), interconnected through adjustible weights are called Neural Networks.They are called Neural networks because the non-linear elements have as their inputs a weighted sum of the outputs of other elements-much like netwoks of biological neurons do. Back propagation proceeds by comparing the output of the network to that expected, and computing an error measure based on sum of square differences.

4.5 Results

Result1:With all features, all classes.. Result2:With only shapecolor features Result3:With only texture features

Class	A	B	C	D	E	Errors
A	7	0	0	0	0	0
B	0	6	1	0	0	1
C	0	0	0	8	0	0
D	0	0	3	0	1	1
E	0	0	0	4	4	4
Errors	0	0	1	4	1	6

Table 4.3: Confusion Matrix All features

SVM Classifer Results

Class	A	B	C	D	E	Errors
A	1	1	1	4	0	6
B	0	7	0	0	0	0
C	4	0	3	0	1	5
D	1	0	0	0	3	4
E	0	1	1	0	6	2
Errors	5	2	2	4	4	17

Table 4.4: Texture features

Class	A	B	C	D	E	Errors
A	7	0	0	0	0	0
B	0	7	0	0	0	0
C	0	0	8	0	0	0
D	0	0	0	2	2	2
E	0	0	1	0	7	1
Errors	0	0	1	0	2	3

Table 4.5: Shape features

Class	A	B	C	D	E	Errors
A	7	0	0	0	0	0
B	0	7	0	0	0	0
C	0	0	8	0	0	0
D	0	0	0	3	1	1
E	0	0	1	0	7	1
Errors	0	0	1	0	1	2

Table 4.6: Shape and Texture features

4.6 Conclusion

Bibliography

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

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

Cseke, I. (1992). 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.

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

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

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

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

Weiss, Y. (1996). Motion segmentation using EM- a short tutorial. Technical report, MIT, MA 02139, USA. E10-120.