MICROSCOPIC IMAGE ANALYSIS USING COLOR TEXTURE CLASSIFICATION

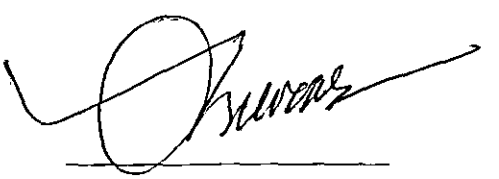
Farhan Dawood
Reg. No. 61-FET/MSEE/F07

Submitted in partial fulfillment of the requirement for the
Master (MS) degree in Electronic Engineering
with specialization in Signal and Image Processing
at the faculty of Engineering and Technology
International Islamic University,
Islamabad.

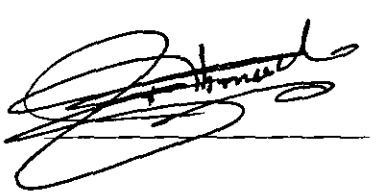
CERTIFICATE OF APPROVAL

It is certified that we have read the project report submitted by **Farhan Dawood (61-FET/MSEE/F07)**. It is our judgment that this report is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for MS degree in Electronic Engineering.

Supervisor
Dr. Tanweer Ahmed Cheema
Assistant Professor,
Department of Electronic Engineering,
Faculty of Engineering & Technology,
International Islamic University, Islamabad



External Examiner
Dr. Anwar Manzar
Associate Professor
AWC (NESCOM)



Internal Examiner
Dr. Bashir Ahmed Zahir
Assistant Professor
Faculty of Engineering & Technology
International Islamic University, Islamabad

DECLARATION

I, hereby declare that this research work has not been copied from any source. It is further declared that we have developed this research work entirely based on our personal efforts under the supervision of Dr. Tanweer Ahmed Cheema. If this thesis is proved to be copied or reported at any stage, we accept the responsibility to face the subsequent consequences. No part of this work inscribed in this thesis has either been submitted to any other university for the award of degree / qualification.

Farhan Dawood

61-FET/MSEE/F07

ACKNOWLEDGEMENT

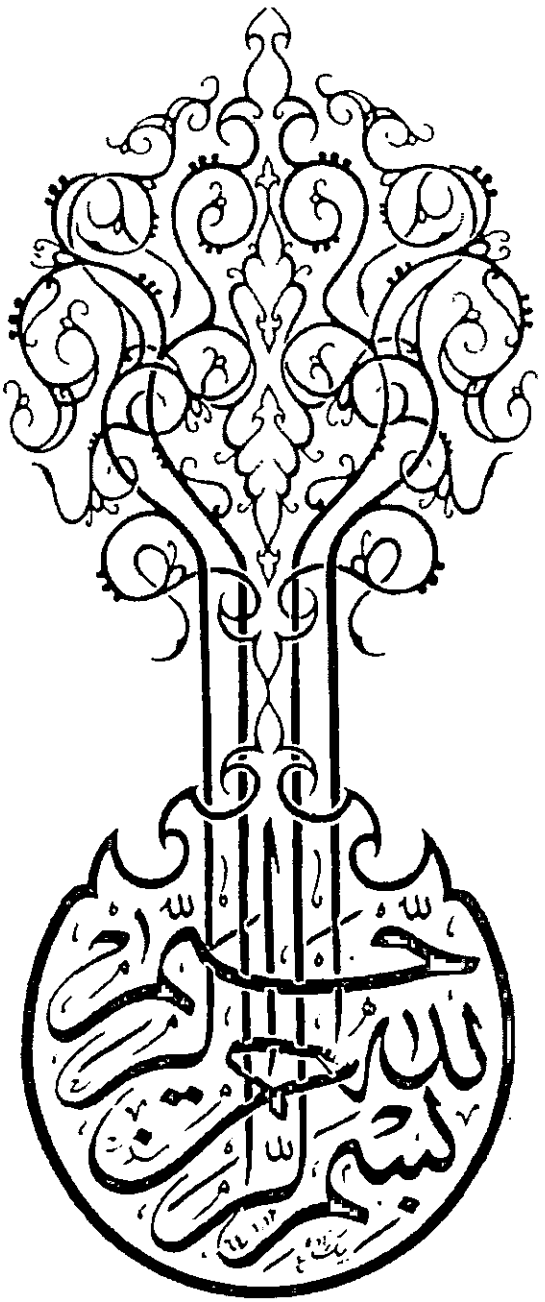
I am heartily thankful to my supervisor, Dr. Tanweer Ahmed Cheema, whose guidance, encouragement and support from the initial to the final level enabled me to develop an understanding of the subject. Throughout my thesis-writing period, he provided sound advice, good teaching, and lots of good ideas. I would have been lost without him.

I am indebted to my class fellows and friends for providing a stimulating and fun environment in which to learn and grow. Furthermore I am deeply indebted to my class fellows that have provided the environment for sharing their experiences about the problem issues involved as well as participated in developing solutions to the identified problems.

Finally I want to thank my family. The encouragement and support is a powerful source of inspiration and energy. A special thank is devoted to my parents for a never ending support.

Farhan Dawood

61-FET/MSEE/F07


بسم الله الرحمن الرحيم

ABSTRACT

Microscopic Image Analysis is getting more exposure in clinical medicine due to the advancement in image acquisition and computer technologies. It is very significant in medical research to specify the correct type of the disease for its accurate treatment. Follicular Lymphoma is a cancer affecting the immune system. World Health Organization divides Follicular Lymphoma into three histologic grades. There are high rates of divergence in analysis of these grades. This results in inter-observer and intra-observer variability when pathologists view the same case multiple times. A two stage system is proposed for the discrimination of Follicular Lymphoma histologic grades to eliminate inter and intra reader variability and maximize consistency of results. In the first stage we differentiate high grades from low grades. For this we have proposed a ratio criterion algorithm. In the second stage we will eliminate the overlapping between lower grades (grade 1 and grade 2). We have proposed a classification mechanism, where cooccurence features are computed from a color approximated/quantized image. Due to the uniform color spectrum of the H&E stained Follicular Lymphoma images we have used neural network based color approximation/quantization approach using Self Organizing Feature Map (SOFM). The proposed system is implemented and tested for various H&E stained FL tissue samples. The first phase classifies the higher grades from lower grades accurately but there is an overlap in the lower grades. This overlapping is removed using the second stage where we have achieved average classification accuracy of 85.7% using k-NN classifier.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Microscopic image processing and analysis of tissue samples with computer vision and image analysis has been an active research area for years. Histopathology is the examination of tissue samples under the high power microscope to have the clear evidence of the disease. Microscopic imaging has been getting a lot more exposure especially in the filed of clinical medicine. It is the most important tool of the pathologist used in clinical diagnosis of diseases such as cancer. The gradual advancement in whole slide tissue imaging technology leads to a vast variety of interpretation and visualization techniques of microscopic images.

Exact type or cause of patient disease must be determined for the exact treatments. Therefore, the exact condition of the patient's disease has to be known. There are high rates of divergence over the patient's condition among pathologists [1]. The percentage of poorly detecting the type of the lesions seems greatest in the diagnosis of cancer, and some studies. This leads to over and under grading of follicular lymphoma cases which causes life-threatening results in patients with this disease [2], [3]. The use of computers to assist the pathologist is growing. The computer assisted

system will make histopathological diagnosis of follicular lymphoma more accurate and consistent, which results in effective treatment of patients.

The proposed technique has wide range of application and can be used for various other medical diagnosis problems. The method is proposed for the classification of follicular lymphoma but it can produce satisfactory results for other diseases such as neuroblastoma classification and assessing the malignancy level of urinary bladder carcinomas.

1.1 Motivation

Lymphoma is a cancer that involves a type of white blood cells named "lymphocytes" [3]. Lymphocyte is the cell in the immune system and is the implicit section of the "lymphatic system" [4]. Lymphoma often develops in the lymph nodes, where it appears as an enlargement of the node and is treated as a tumor. Lymphoma, the most common type blood cancer, occurs when lymphocytes grow abnormally. The two main types of lymphoma are NHL and HL. The most common of the slow growing type of NHL is Follicular Lymphoma. The lymph nodes affected by this lymphoma show a round shaped structure when viewed under the microscope [2], [3], [4]. These round structures are called "follicles". Therefore, this type of lymphoma is called "follicular lymphoma" [4].

Follicular lymphoma is diagnosed using biopsy of the lymph node. A small sample of the diseased node is taken and examined under the microscope by a pathologist.

Grading determines the nature of the lymphoma. Huge variations occur among the results provided by the pathologists because of the observer's capability to view it.

The divergence among observers is increasing and causes various problems for the patients. Therefore, it has been decided to non-subjectively and quantifiably analyze the rapid cellular growth in FL, along with the current grading systems. The development of a computer-assisted grading system of follicular lymphoma has been proposed in this literature. This system will assist pathologists to accurately classify the grade of this disease.

1.1.1 Grading Criteria

Follicular lymphoma is graded by the ratio of "centroblasts". "Centroblasts are non-cleaved follicular centre cells". World Health Organization [4] recommends a three-grade system (Grade I, II & III), by counting the amount of centroblasts. "Grade I cases have 0-5 centroblasts/HPF; Grade II cases have 6-15 centroblasts/HPF; Grade III cases have >15 centroblasts/HPF" [4]. Generally, the level of maximum magnification used as HPF is 400x magnification level.

According to the WHO criterion [4], the disease is graded into the following three grades. The Grade I follicular lymphoma contains the monotonous portions of small lymphoid cells with irregular, angulated nuclei, dispersed chromatin (located in cell nuclei); nucleoli are not prominent or easily noticeable, and scant cytoplasm [4]. The Grade II FL contains the mixture of different sizes of lymphoid cells. Generally the higher grade sample of follicular lymphoma act aggressively, therefore the most

significant distinction is between the grade 1 and 2 cases (lower grades) and, grade 3 cases (higher grades).

1.2 Problem Statement

The problem is in diagnosis of exact grade of the Follicular Lymphoma. We propose a system that will help the diagnosticians in the grading of follicular lymphoma. The purpose is to develop and design techniques and methods for extracting features for classification.

In this study, we propose a two stage classification system that classifies images into their gibing grades based on standard criteria discussed above. The first stage differentiates the lower grades (grade I and II) from the higher grades (grade III) using the features extracted through segmentation [31]. The second stage classifies the images into grades 1 and II based on the statistical features extracted from the color quantized images of FL. Since the lower grades are difficult to differentiate so we proposed a low level processing using non-linear color quantization using SOFM. This technique will help us to properly discriminate grade 1 from grade 2.

For differentiating higher grades (grade 3) from lower and intermediate grades (grade 1 and 2),

- Conversion of image to the La*b* space for better processing. The reason for this conversion is discussed later.

- The colored image is then segmented into desired components using the k-means.

- For discriminating grade 3 from grade 1 and 2 we proposed a ratio criterion.

For further differentiating intermediate and lower grades (grade 1 & 2),

- Quantize the color image using the Self Organized Feature Maps Algorithm.

- Calculate the co-occurrence matrix from the color approximated/quantized image.

- Extract co-occurrence features namely homogeneity, correlation, contrast, energy and entropy.

For classification the results are passed on to the classifier. We consider using k-nearest neighbor method for this purpose. Other classification techniques such as Bayesian classifier are also tested on the proposed method for grading prognosis.

1.3 Outline of Thesis

This dissertation consists of four chapters. The next chapters will explain the proposed methods used for classification.

Chapter 2 – Overview of Computing Tools: This chapter gives a brief overview of the methods used for the proposed system. The use of segmentation is explained and the k-means segmentation algorithm is used. Next we provide the overview of the features extracted for classification for stage 1.

Chapter 3 – Feature Classification using Color Texture Analysis: In chapter 3 we have proposed the method for differentiating between grade I and II. For this purpose we have proposed a method for extracting the cooccurence matrix features from the H&E stained image using color quantized images. The quantization is done through Kohonen Self Organizing feature Maps.

Chapter 4 – Simulation Results: The experimental and simulation results of the suggested method are discussed in this chapter.

Chapter 5 – Conclusion and Future Directions: The results are summarized in this chapter and the main conclusions of this dissertation and possible future directions are discussed.

CHAPTER 2

OVERVIEW OF COMPUTING TOOLS

In this chapter, we will explain the methodology involved in classification of the grades of Follicular Lymphoma images and provides an overview of the computing tools used in classification of these images. The unsupervised segmentation methods have been used to extract regions of interest and perform quantitative analysis on them. The novel technique of color quantization using neural network has been and extract statistical features are extracted for further accurate classification of the images for different grades.

"Image analysis in clinical pathology" [5] has wide range of applications. The every day increasing capability of computers allowed for more complicated examination and evaluations of cell and tissue characteristics. Various systematic methods [6] were proposed by the researchers towards the advancement of histopathological image analysis especially FL images. A hybrid technique has been used for the classification of follicular lymphoma [7]. They developed a hybrid approach that combines information from several slides with different techniques of stains.

2.1 Image Segmentation

"Image Segmentation" is the process of dividing the image into different sections based on the components in the image. This will help the user to analyze the system more efficiently and evaluate it by parts. These segmented images are analyzed easily and efficiently.

2.1.1 Color Space Conversion

The RGB or Red-Green-Blue color space has been commonly used in image processing because of the information available by camera apparatus. RGB, however, is not a uniform color space [10]. Color based segmentation is greatly influenced by the selection of color space. The L*a*b* is a better representation of content of the colors in an image. Hence, this color space is an ideal choice for segmentation.

The L*a*b* is a three argument space with dimensions L for lightness, a and b for the color-opponent dimensions. The major advantage of this space is that "distance metric used for segmentation techniques is Euclidean" [10]. The L*a*b* is based on an intermediate system, known as the "CIE XYZ space", which is derived from RGB as follows [11]:

$$X = 0.412\,R + 0.357\,G + 0.180\,B \qquad 2.1$$

$$Y = 0.212\,R + 0.715\,G + 0.072\,B \qquad 2.2$$

$$Z = 0.019\,R + 0.119\,G + 0.950\,B \qquad 2.3$$

Based on this definition, L*a*b* is defined as follows:

$$L^* = 116\, f\left(\frac{Y}{Y_n}\right) - 16 \qquad 2.4$$

$$a^* = 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right] \qquad 2.5$$

$$b^* = 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right] \qquad 2.6$$

Where

$$f(q) = \begin{cases} q^{1/3} & if\ q > 0.008856 \\ 7.789q + \frac{16}{116} & otherwise \end{cases} \qquad 2.7$$

X_n, Y_n *and* Z_n represent white as defined by the CIE standard *illuminant*.

The L*a*b* requires a conversion to another space, but still its range encompasses the entire visible spectrum and can represent accurately the colors of any display. This feature makes this color space useful in both image manipulation and image compression applications [12].

2.1.2 K-Means Clustering Algorithm

The K-means is an unsupervised clustering algorithm in image processing. K-means is commonly used in computer methods vision as a tool for sectionalization. It is also used for selecting useful objects from the image for further processing and operations.

Basic Algorithm

K-means algorithm is initialized by defining k-centroids, one for each cluster. It then creates a partition by taking each data point of a given data set and associates it with

the closest centroid. The algorithm is repeated and the k new centroids recalculated for new clusters. A loop has been generated which results in a new binding between the same data points. The convergence of algorithm is obtained when the points no longer change their locations.

The target is to minimize "intra-cluster variance", or the "squared error function"

$$E = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \qquad 2.8$$

Where there are k clusters, $S_i = 1,2,\ldots,k$ and μ_i is the centroid or mean point of all the points $x_i \in S_i$.

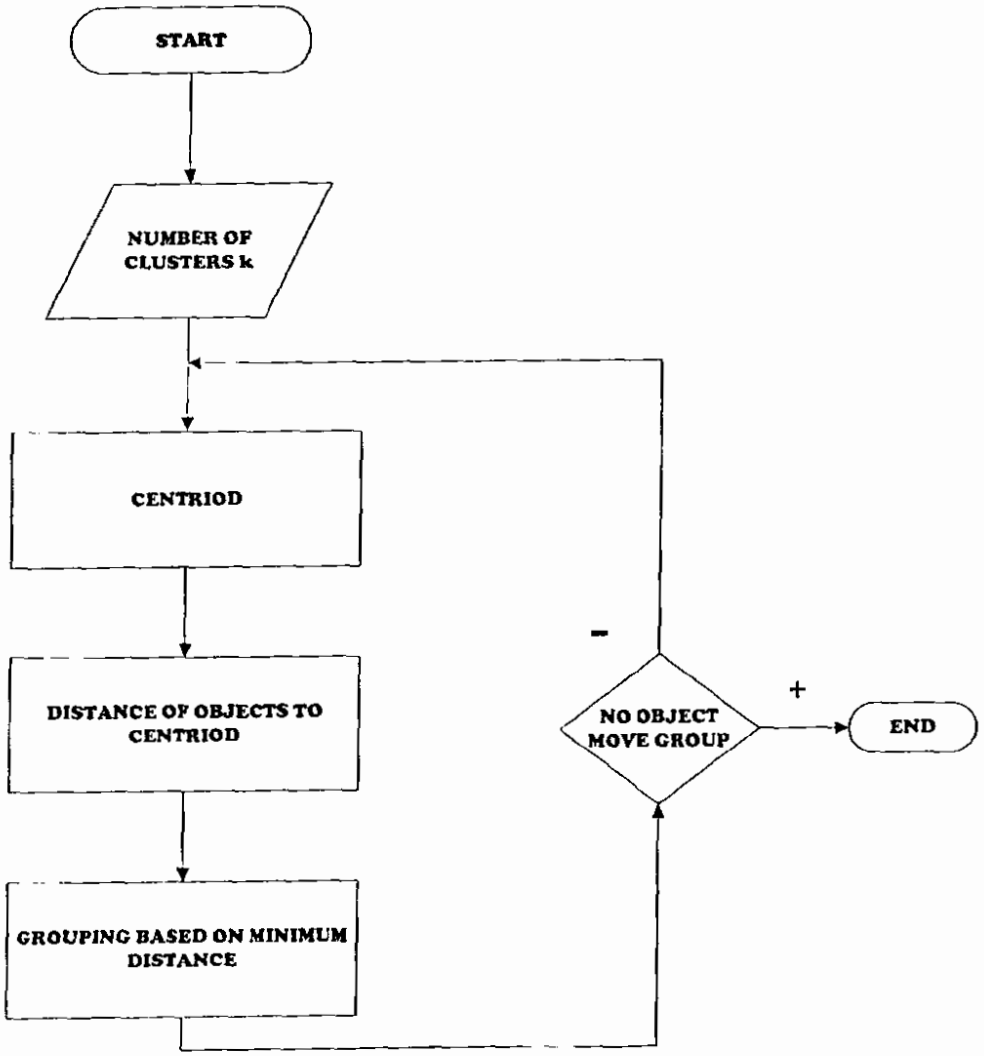


Figure 2.1 – Flow chart of k-means algorithm

2.2 Color Quantization

"Color quantization" or color approximation is the procedure to minimize the number of colors that are used for limited number of times. These distinct colors are replaced by the approximate values. The huge number of colours involved in images brings about unwanted and undesirable computational expenses [13]. Generally, quantization is the process of selecting L vectors in some D dimensional space to represent D vectors from that space where $L < D$ and the total amount of error induced by the quantization is minimized.

Various color reduction algorithms have been proposed in the literature. The function of compression algorithms is to minimize the colours of an image such that there is an optimum visually sensual similarity between an original color image and the approximated reduced image and thus achieve good results. "The most commonly used methods for reduction of colors in an image are based on nearest colors merging or color error diffusion. A color palette may be fixed or adaptively created. Fixed color palettes have already been produced while the adaptive color palettes are created initially by scanning the image and selecting the minimum colors and then to create the quantized image by mapping the original color points to these selected colors. These techniques are not efficient for image analysis operations". [13], [16], [17], [18]. More recently, unsupervised clustering methods have become useful. These techniques made the use of spatial information which produces visually sensible and good results. The resulting algorithms are quite complex but they are efficiently implementable with better results when compared to other old methods.

2.3 Self Organizing Feature Maps (SOFM)

2.3.1 Overview

The self-organizing Map (SOM) is effective tool for the visualization of high-dimensional data. The aim of the self organizing maps is to transmute input pattern of discretional dimension into a one- or two- dimensional distinct map. The nerve cells turned towards an organized and meaningful form based on the patterns provided by the input samples through the process known as "competitive learning" [19]. A self-organizing is therefore described by the formation of a topographic map of the input patterns [19], [20].

The SOFM consists of the input and the output layers. An example of a planar array of neurons with hexagonal neighborhoods is shown in Figure 3.

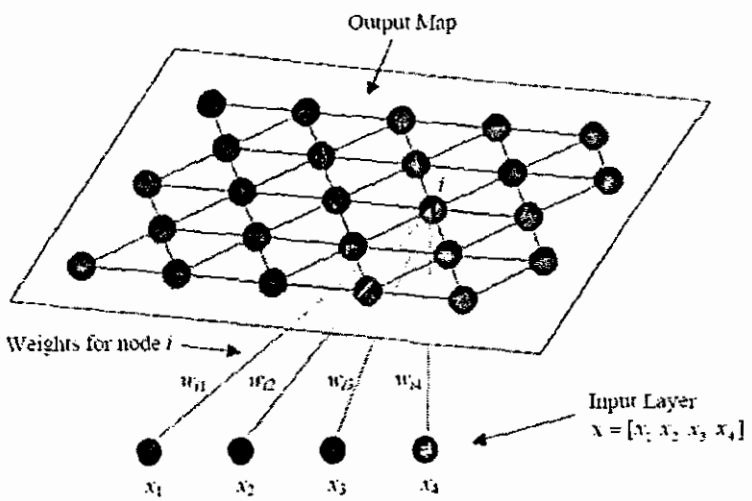


Figure 2.2: Example of mapping input vector **X** to SOFM. The input vector is connected to the output Map via the weight vector **W**.

2.3.2 Kohonen Learning Algorithm

The function of the learning algorithm is to made different nodes in the lattice to react effectively to the input samples. This algorithm is processed first by initializing the weight vectors in network. The weights are initialized by randomly assigning values to them. After initialization the following essential steps are processed [21]:

- **Competitive Process**

At the application of some input pattern the neurons in the lattice tends to compete among themselves to become activated to specific input values. The activated neuron is called the winning neuron and this process is called competitive process.

The input pattern randomly from the space of dimension m denoted by

$$x = [x_1, x_2, \ldots\ldots\ldots, x_m]^T \qquad 2.9$$

The dimension of the weight vectors must be equal to the dimensions of the input pattern. Suppose the weight samples be,

$$w_j = [w_{j1}, w_{j2}, \ldots\ldots\ldots, w_{jm}]^T \ , \ j = 1, 2, \ldots\ldots, l \qquad 2.10$$

The weight vector with the shortest distance is the winner. Let $i(x)$ be the neuron that matches the input,

$$i(x) = arg\ \min_j \|x - w_j\| \ , \ j = 1, 2, \ldots\ldots., l \qquad 2.11$$

The neuron which satisfies the above constraint is the "best-matching or winning neuron" [19], [20].

In the learning process the nodes in the lattice that are closed to each other at a certain distance will activate each other with the application of same input pattern. The winning neuron will tend to activate the neighboring neurons. This results in the ordering of the overall neurons in the lattice. The input neuron tends to activate the neurons in its immediate neighborhood. As the learning continues the neighborhood area shrinks gradually [19].

- **Adaptive Process**

In self organized (SOFM) network, the weight w_j of neuron j in the network is required to change comparatively to the input x [18]. The updating process in the discrete-time notation, given the weight vector $w_j(n)$ jth neuron at time n, the update vector $w_j(n+1)$ at time $n+1$ is defined by

$$w_j(n+1) = w_j(n) + h_{j,i(x)}(n)\left(x - w_j(n)\right) \qquad 2.12$$

The Adaptive Process is employed to the neighboring neurons of the excited neurons. The learning rate, $\eta(n)$ decay with time.

2.4 Feature Extraction

Feature Extraction is the transformation of the input data into set of features. Numerous methods that were designed for specific application have been proposed by different researchers and authors. However, there seems to be no general method or a formal approach which is useful in a broad range of images. Gray level cooccurence matrices carry a significant amount of texture information in images and perform better than many other texture extraction algorithms. This seems to be a good choice for our application.

The initial work on texture discrimination using gray level spatial dependencies was done in early seventies. In some comparative studies, researchers have observed that GLCM matrices were very successful in discriminating images with relatively homogeneous textures. A comparative study was done in [21] used Markov-generated images to evaluate the performances of different texture analysis algorithms for automatic texture discrimination. One of the early approaches that use spatial relationships of gray levels in texture discrimination is [22], where Haralick used features were computed from the co-occurrence matrices for automatic scene identification of remote sensing images and achieved 70% accuracy.

From these experiments, it can be concluded that spatial gray level dependency matrices carry a significant amount of texture information in images with some homogeneously textured regions and perform better than many other texture extraction algorithms that were listed above. This seems to be a good choice for our application.

2.4.1 Gray Level Co-occurrence

The "Grey Level Cooccurence Matrix", GLCM [23] is defined as a "sample approximation of the joint PDF of the gray levels of two pixels separated by a given displacement" [23], [31]. GLCM is a matrix containing the values explaining the amount of times a specific arrangement of the pixels co occur together given a specific distance and orientation of the group of pixels under consideration [23]. The joint distribution of the pair of pixels separated by some distance and at specific orientation is denoted as $P(i, j; d, \theta)$.

For the arrangements of the pixel in the image shown in Figure 2.3, gray level co-occurrence matrices can be defined as

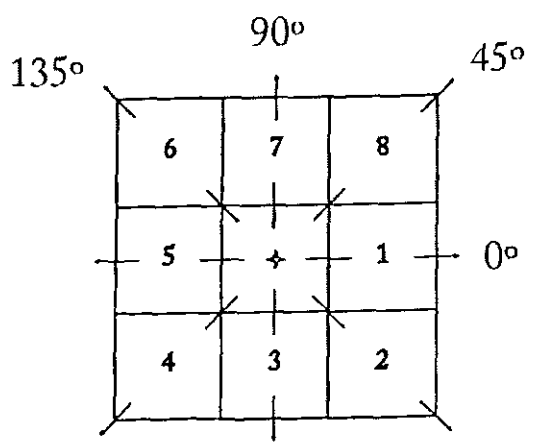


Figure 2.3: Spatial Arrangements of Pixels

Formally, the orientation θ is in four 0°, 45°, 90°, and 135°. For each combination d and φ, a two dimensional histogram is defined,

$$P(i,j;d,0°) = \#\left\{\begin{array}{l}((k,l),(m,n)) \in (L_x \times L_y) \times (L_x \times L_y) \mid \\ k-m=0, |l-n| = d, I(k,l) = i, I(m,n) = j\end{array}\right\} \quad 2.13$$

$$P(i,j;d,45°) = \#\left\{\begin{array}{l}((k,l),(m,n)) \in (L_x \times L_y) \times (L_x \times L_y) \mid \\ k-m=-d, |l-n| = d, I(k,l) = i, I(m,n) = j\end{array}\right\} \quad 2.14$$

$$P(i,j;d,90°) = \#\left\{\begin{array}{l}((k,l),(m,n)) \in (L_x \times L_y) \times (L_x \times L_y) \mid \\ k-m=d, |l-n| = 0, I(k,l) = i, I(m,n) = j\end{array}\right\} \quad 2.15$$

$$P(i,j;d,135°) = \#\left\{\begin{array}{l}((k,l),(m,n)) \in (L_x \times L_y) \times (L_x \times L_y) \mid \\ k-m=d, |l-n| = d, I(k,l) = i, I(m,n) = j\end{array}\right\} \quad 2.16$$

Here # denotes the total components in the typeset. All the matrices computed are symmetric i.e. $P(i,j;d,\theta) = P(j,i;d,\theta)$.

Figure 2.4 shows an example for co-occurrence matrix computation from [23] for an image with four gray level values 0, 1, 2 and 3.

2.4.2 Haralick Features

Haralick computed fourteen features exploiting the properties of the co-occurrence matrices [23]. As we have seen earlier in section that with different values of distance and directions of the pair of pixels we can generate different cooccurence matrices. Therefore, using this property we can create a large feature set for these different matrices. We decided to use only the five of these features namely Homogeneity, Entropy, Energy, Contrast and Correlation.

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

(a)

		Gray Level			
		0	1	2	3
Gray Level	0	#(0,0)	#(0,1)	#(0,2)	#(0,3)
	1	#(1,0)	#(1,1)	#(1,2)	#(1,3)
	2	#(2,0)	#(2,1)	#(2,2)	#(2,3)
	3	#(3,0)	#(3,1)	#(3,2)	#(3,3)

(b)

$$P(i,j;1,0°) = \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

(c) $(d,\theta) = (1,0°)$

$$P(i,j;1,45°) = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix}$$

(d) $(d,\theta) = (1,45°)$

$$P(i,j;1,90°) = \begin{pmatrix} 6 & 0 & 2 & 0 \\ 0 & 4 & 2 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix}$$

(e) $(d,\theta) = (1,90°)$

$$P(i,j;1,135°) = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

(f) $(d,\theta) = (1,135°)$

Figure 2.4: (a) Image with gray level values. (b) Possible GLCM orientation for image (c)- (f) Computation of all four matrices at d=1 and given directions.

2.5 Feature Classification

"K-Nearest Neighbor" is a method for the classification of different data sets on the basis of nearest distributions in the given set [29]. The selected data point is assigned to group with most common data points amongst its k nearest neighbors. The neighbors are selected during the training phase. In training step, the neighbors are selected for which the correct classification of the data set is known. The regions are divided in the training space by locations and labels. The training phase only defines

and stores the labels of the data set. When the test data is applied to already trained set, distance from the new data point is measured to all stored labels and k closest samples are selected [30].

The value of k is very crucial for classification and depends upon the data used. Very large values of k may reduce the effect of noise during classification but may lead to less distinct boundaries between classes. The stepwise description of the algorithm is as follows: Suppose we have a data set S_D which represents classes C. Suppose we want to assign a particular data point a to one of these C classes.

- Define suitable k.
- Find k nearest to the test point a, let them to be $n_1 \cdots n_k$.
- Make a hyper-cycle d_h around a with radius n.
- Classify a as a class C which has the maximum votes in plane d_h.

CHAPTER 3

FEATURE EXTRACTION AND CLASSIFICATION

Computerized image analysis provides precise predictive clews about the disease not easily observed by analysis performed by doctors. The computerized system prevents the variability in reproducing the data. In this chapter the complete proposed methodology is explained in detail.

As discussed earlier in chapter 1 that the study is divided in two stages. In first stage the features were extracted from the microscopy images. We have used the method of k-means segmentation which was discussed in detail in chapter 2. On the basis of these features the images are classified as lower or higher grades. In the next stage we have extracted another set of features using color quantization and cooccurence matrix. These features are useful for the classification between lower grades. Figure 3.1 gives the graphical overview of the proposed system used for the grading of follicular lymphoma.

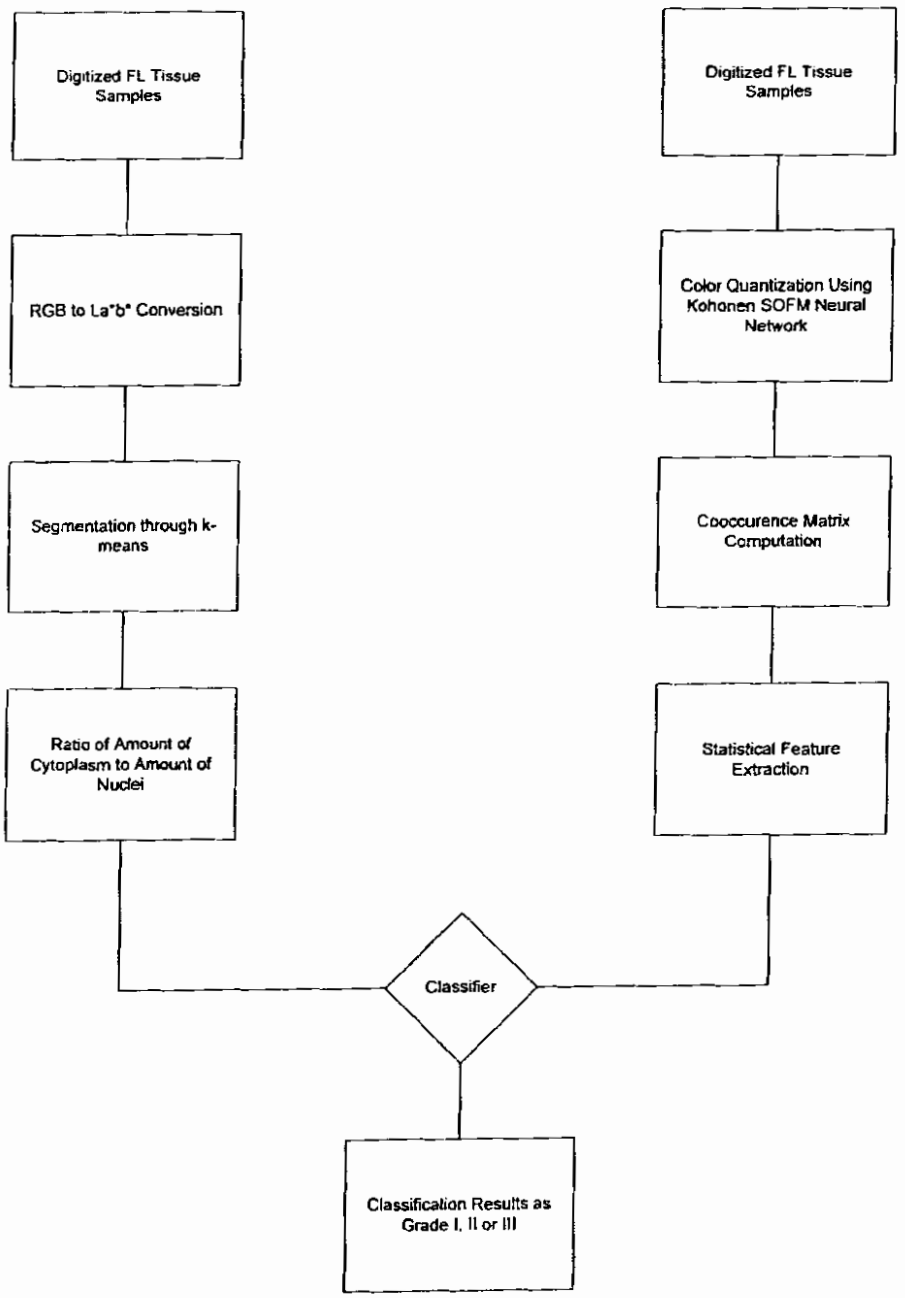


Figure 3.1 –Diagram of the proposed Image analysis and classification system for the grading of Follicular Lymphoma

3.1 Differentiating Low Grades from High Grades

As we know that all the three grades have different arrangements of cytological components which differs them from others but they are difficult to examine and analyze using microscope. In addition to these components, there are also background regions that do not correspond to any tissue component [31]. Using these details of the images we performed partitioning using K-means algorithm to distinguish nuclei, cytoplasm and extracellular material, and Threshold segmentation to extract RBC and background.

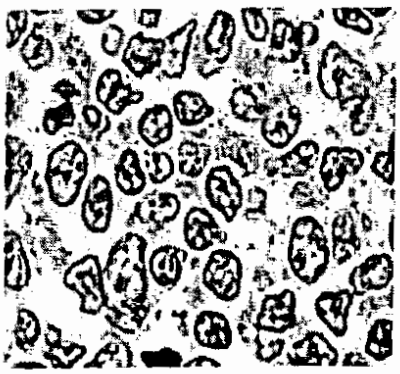
We performed image segmentation in the L*a*b* space using the K-means. The value of k is initialized as three, which represents "nuclei, cytoplasm and extracellular material". As discussed earlier in this chapter that RBCs and background regions are not consistent with any other tissue component. The threshold segmentation is done to extract the RBC and background from the FL images. After threshold segmentation these regions are omitted from the image before k-means clustering is applied [31].

The threshold segmentation for RBC is:

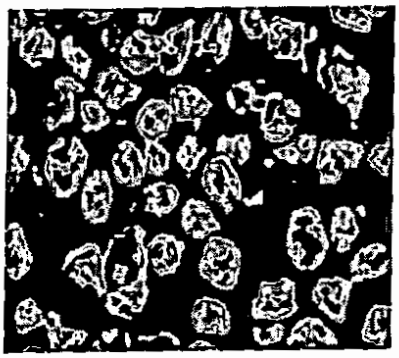
$$RBC = \begin{cases} 1 & if \ \frac{r(i,j)}{r(i,j)+g(i,j)+b(i,j)} \geq \alpha \\ 0 & otherwise \end{cases} \quad 3.1$$

Similarly, for the background region:

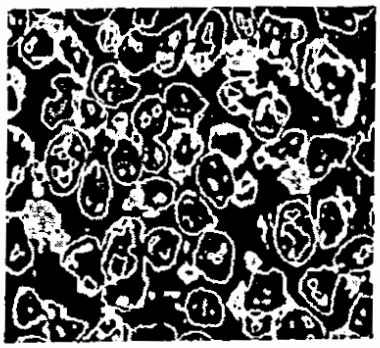
$$BG = \begin{cases} 1 & if \ (r(i,j) \geq) \& (g(i,j) \geq \beta) \& (b(i,j) \geq \beta) \\ 0 & otherwise \end{cases} \quad 3.2$$



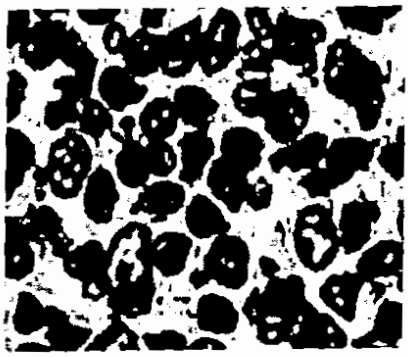
(a) H&E stained FL image.



(b) Nuclei in the image.

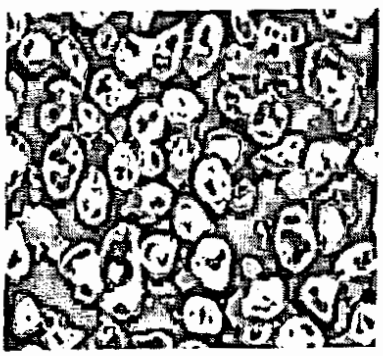


(c) Cytoplasm in the image.



(d) Extracellular Materials.

(e) Image labeled by cluster Index.



(f) Pseudo-colored clustered image.

Figure 3.2: Segmented images using k-means.

3.1.1 Ratio Criterion

"There are larger numbers of centroblasts in higher grades follicular lymphoma samples, therefore the, images of higher histological grades have less homogeneous organization of nuclei and cytoplasm components with respect to their relative spatial distributions when compared to the lower grade samples in which those regions are more compact and evenly distributed" [4].

Mathematically this observation from the follicular lymphoma images based on our segmentation method as follows [31]:

$$\zeta = \frac{\sum_{i,j} \chi}{\sum_{i,j} \upsilon} \qquad 3.3$$

where i, j are the pixel locations, χ is the cytoplasm and υ is the nuclei pixels.

On the basis of this observation, it is clearly notified that the lower grades show large variation from the higher grades, but there is some overlap in the lower grades.

3.2 Color Quantization Using SOF Map

For this purpose, a color quantization scheme has been proposed which operates on the principle of neural networks. We have used Kohonen Self-Organizing Maps for this purpose.

The main goal of a SOFM neural network is the reduction of a large set of input vectors with a smaller set of vectors, so that a 'good' approximation of the original

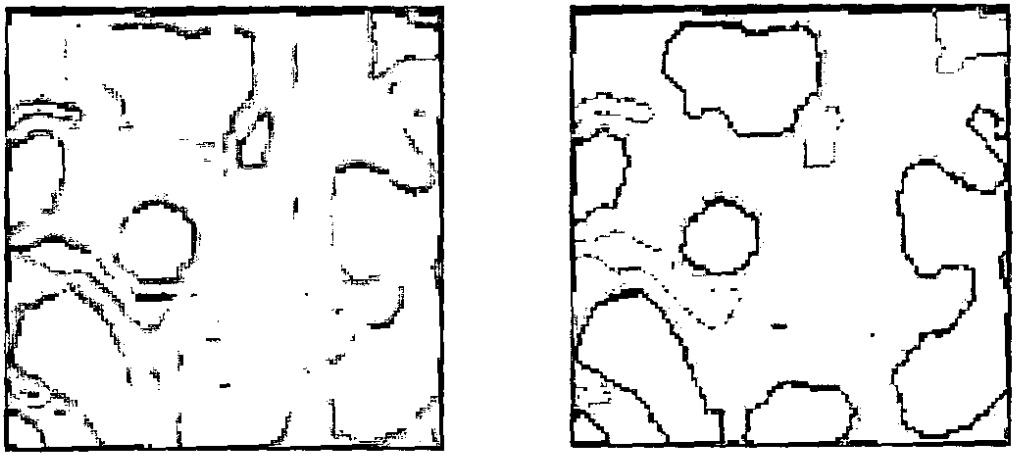
input space to be obtained. In other words, a SOFM neural network decreases the input feature space into a proper smaller one. The resultant feature space can be viewed as a representative of the original feature space, and therefore, it satisfies the main statistical characteristics of the input space.

After training, the optimal RGB values are obtained by the neural network. Next, the original image is rescanned and by using the neural network, new image is constructed with less number of colors.

The algorithm for color quantization is shown below:

1. *Choose total number of colors required after quantization (e.g. 8 colors).*
2. *Initialize those colors by random values.*
3. *For count = 1 to 500*
   a. *Take first pixel of given image and calculate the color difference of that pixel with required colors*
   $$d = |R - R'| + |G - G'| + |B - B'|$$
   b. *Choose the minimum distance as winner*
   c. *Update the color in the table using*
   $$R'(n+1) = R'(t) + \eta(t)(R - R')$$
   $$G'(n+1) = G'(t) + \eta(t)(G - G')$$
   $$B'(n+1) = B(t) + \eta(t)(B - B')$$
   where R', G', B' are Red, Blue and Green colors in required table, and RGB are actual colors of corresponding pixel, and $\eta(n)$is the learning rate parameter.
   d. *Do steps (a) to (c) for all pixels in the image.*
4. *Reconstruct the approximate image by using optimum colors in the table.*

The algorithm stops execution after a predefined number of iterations or until there is a substantial change in the image.



(a) Pseudo-colored Image (b) 16-Color Quantized Image

(c) 8-Color Quantized Image (d) 32-Color Quantized Image

Figure 3.3: Color Quantization Using SOFM.

3.2.1 GLCM Features for Color Images

We have extracted features from the colored images. Approaches used are based on existing grey level methods that are adapted into account the color information. In the literature, three families of approaches can be found [28].

The first consists in the use of joint color-texture features. Textural features and color features are computed individually and then are used together as the entry of a classifier. The second consists in the reduction of the color information in the color images using a quantization method. The images obtained are coded like a grey scale image, each grey scale corresponding to a different color. So both approaches use a transformation of the color images to be able to apply existing grayscale methods, without giving a new definition to color texture. The last consists in taking into account the correlation between the color bands while computing the texture features. In this research work we exploited such a multispectral method [29] considering the correlations between the color bands. This multispectral method seems to be a very good approach for color texture classification problems, in comparison with the other approaches.

The features computed within color bands (without correlations) and between colors bands (correlations only) show that both are complementary. The correlations between color bands introduce new and relevant information. The features are computed directly from the RGB color space. The GLCM matrices are computed for every channel. In this case, color images are coded on three channels, leading to six

different matrices: $(R,R),(G,G),(B,B)$ that are the same as gray scale cooccurence matrices computed on one channel and $(R,G),(R,B),(G,B)$ that take into account the correlations between the channels. So this method led to a total of 30 features [28].

The multispectral method gives very good results for a classification problem. The use of cooccurence matrix method computed both within and between the color bands allows achieving good results.

CHAPTER 4

SIMULATION RESULTS

The performance evaluation of the proposed method for the classification of histopathological grades of follicular lymphoma is discussed in this chapter. The proposed system was explained in detail in chapter 3. The experimental setup is split into two stages. Initially, the images were classified into lower and higher grades using proposed ratio criterion. In the second phase the images are further classified into grade I and grade using color texture analysis.

4.1 Image Data set

The H&E stained whole slide Follicular lymphoma images obtained were digitized using a digitizer at magnification level of $40x$. A total of 510 images were obtained that contains discrete amount of samples of all the histological grades. Out of these 510 images there were 243 samples of grade I, 109 samples of grade II, and 158 samples of grade III. Each of these image samples has a resolution that is equivalent to one microscopic HPF. The training image data set contains 170 samples. The training group have 81 samples of grade I, 36 samples of grade II and 53 samples of

grade III. Similarly, the test image data set contained 162 samples of grade I, 73 samples of grade II and 105 samples of grade III. This makes a total of 340 image samples used in test phase. The constructed dataset of 510 images have a spatial resolution of 2,165 × 1,365 pixels that is equivalent to one microscopic HPF.

4.2 Image Segmentation

The proposed two-stage system is processed as follows. During the first stage, the digitized FL images are used for the classification. The features were extracted using a simple ratio criterion proposed in chapter 3. The histograms are plotted for ratio values of grades after applying ratio criterion. It is clear from the histogram distribution that this method is good for the classification of higher grades and lower grades, but there is overlap between the regions of lower grades i.e., grade I and II.

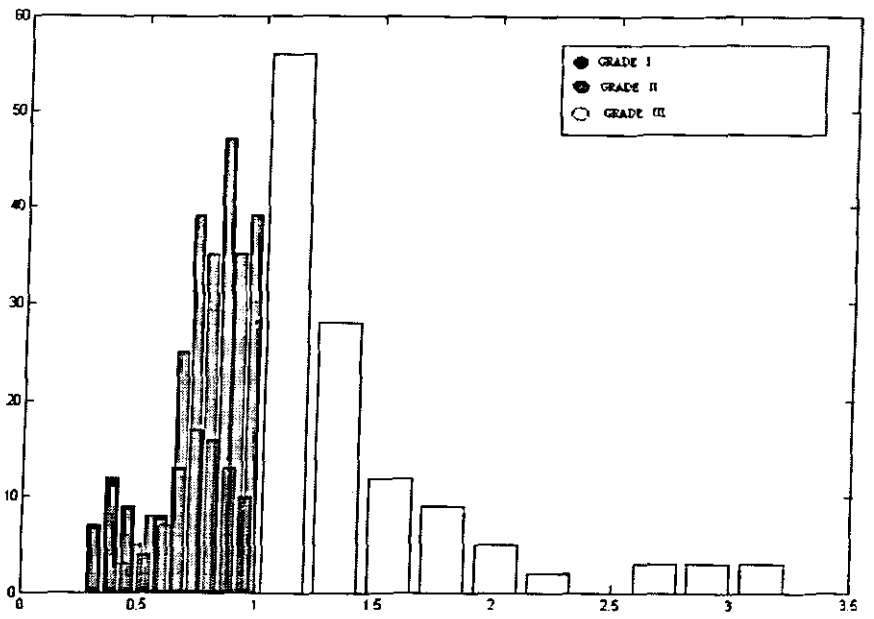


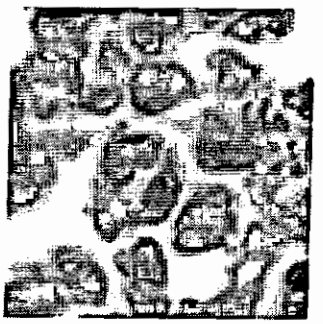
Figure 4.1 – Result of the ratio criterion.

Fig 4.1 demonstrates the results of assortment of images using ratio criterion. It can clearly be seen that grade III images can easily be classified from grade I and grade II, however, there is an overlap between lower grades. The discrimination accuracy has been achieved for the classification among grade I&II and grade III. The overlap between lower grades has been removed using the low level color texture analysis.

4.3 Color Quantization Using Kohonen SOFM

The low level color texture analysis exploit the color spectrum of the H&E stained images. To further classify lower grades we have exploited the technique of color quantization using Kohonen Self Organizing feature Neural Network. We have extracted five features from these color quantized images using Gray-Level Cooccurence Matrix, namely homogeneity, contrast, correlation, entropy and energy. The cooccurence matrix was computed using multispectral method. In multispectral method the correlation was computed for both within color bands (RR, GG, and BB) and between the color bands (RG, RB, BR, BG, GR, GB). Here R is used for Red, G for green and B for Blue. It is observed that the incorporation of color features significantly enhanced the performance of the classifier.

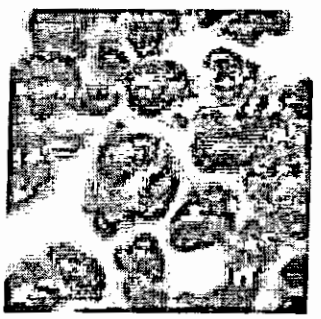
Fig 4.2 shows the images which are quantized using Kohonen Self Organizing Feature Map algorithm. This color reduction method using neural network minimizes the loss of useful information.



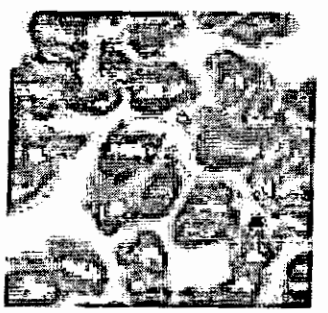
(a) Original Grade-I image.

(b) 8-color Quantized Image

(c) 16-color Quantized Image



(d) 32-color Quantize Image



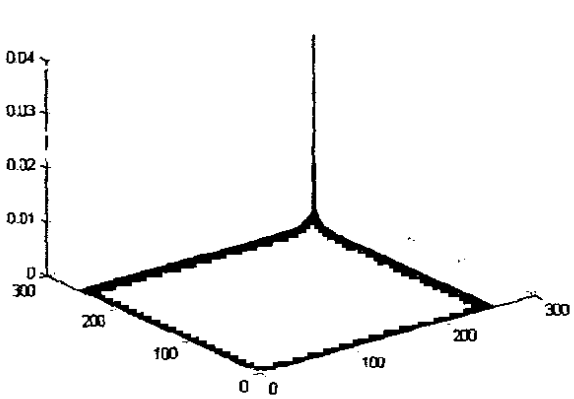
(e) 64-color Quantized Image

Figure 4.2 – Color Quantization using Kohonen Self-Organizing Neural network.

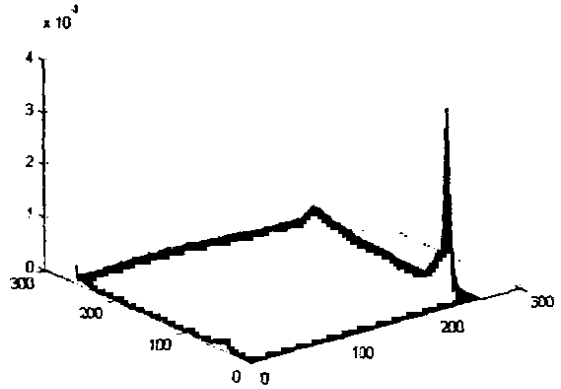
4.4 Cooccurence Matrix Computations

The cooccurence matrix was computed for different color quantization levels. The cooccurence matrix was calculated for the colour images. We have exploited the concept of Grey-Level Cooccurence Matrix (GLCM) and modify it for color images. The features were computed within color bands and between the color bands. For between color bands we have taken the correlation between these color bands. Five Haralick Features were calculated for each computed matrix and the matrix were computed using the correlation between color bands and within color bands. Six matrices were computed for each image making a total of 30 features fro each image. The correlation matrices computed are shown below.

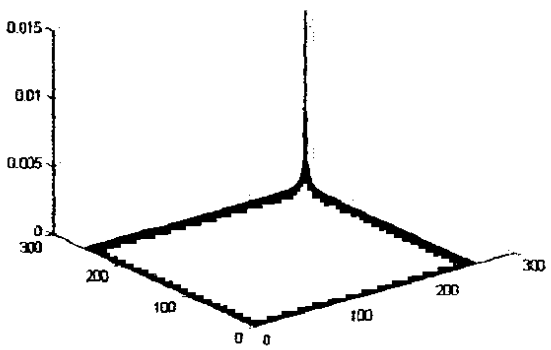
Figure 4.3 shows the six correlation matrices computed between color bands and within color bands. The graphs in Figure 4.3 show the distinct correlation peaks at different points. Every correlation matrix is different from other and will eventually results in distinct features useful in discriminating the histologic grades of Follicular Lymphoma.



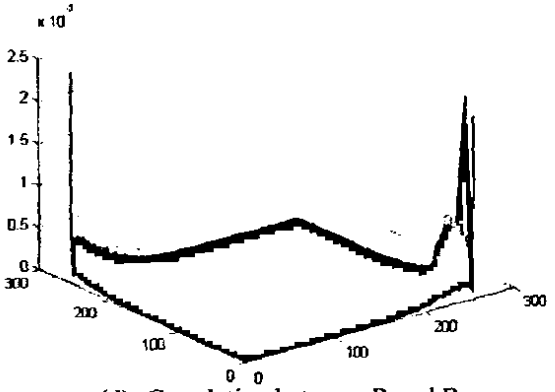
(a) Correlation between B and B



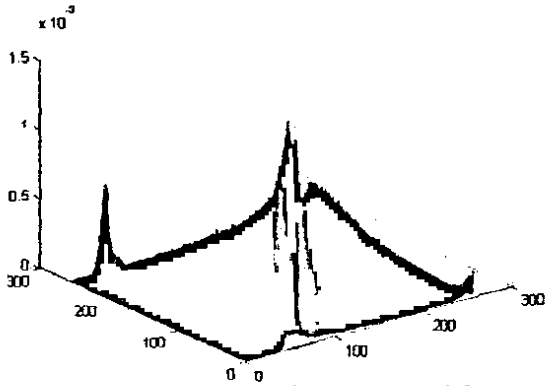
(b) Correlation between G and B



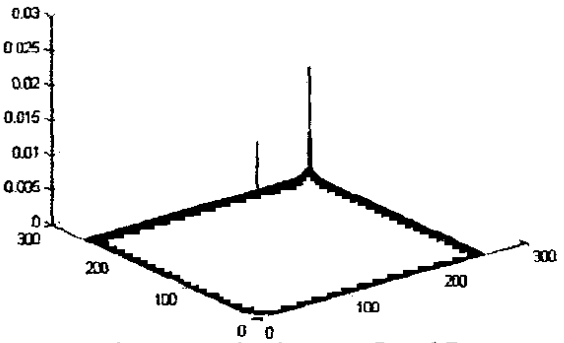
(c) Correlation between G and G



(d) Correlation between R and B



(e) Correlation between R and G



(f) Correlation between R and R

Figure 4.3: Six cooccurence matrices computed using multi-spectral method.

The classification results for Haralick Features were computed using k-nearest neighbour. Table 4.1, 4.2 and 4.3 shows the average percentage of classification rates computed over this feature space using different test and training data sets and with different levels of color quantization. The Training and Testing Data is divided into three groups for evaluating the performance of method for different numbers of data sets.

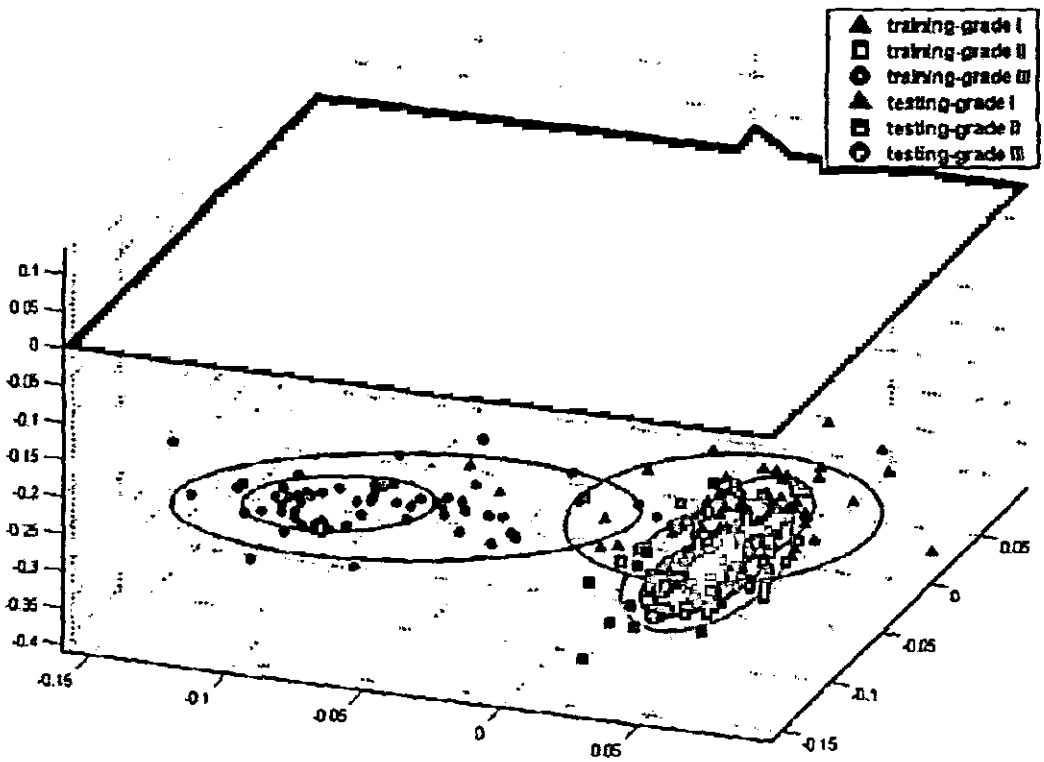


Figure 4.4: Classification using k-NN algorithm.

Figure 4.4 shows that the proposed method is robust enough to classify the images using the extended set of features calculated using the color gray level cooccurence matrix method. This figure shows the decision boundaries estimated using the Bayesian probability method.

Table 4.1: Classification accuracies for discriminating histopathological grades using low-level color texture analysis for different levels of color quantization for Test and Training Data I.

TEST AND TRAINING DATA – I				
COOCCURENCE MATRIX SIZE	**GRADE I**	**GRADE II**	**GRADE III**	**OVERALL CLASIFICATION ACCURACY**
8	94.4 %	66.6 %	88.9 %	96.7 %
16	94.4 %	81.2 %	100 %	98.3 %
32	94.4 %	72.9 %	100 %	97.2 %
64	88.8 %	75.0 %	91.6 %	96.6 %
128	94.4 %	93.7 %	100 %	99.2 %
256	75.5 %	81.2 %	100 %	96.6 %
512	90.5%	88.4%	99.6%	89.5%

Table 4.1 shows the classification accuracy of the features extracted using Haralick features. The k-NN classifier is trained using the training data set I described above. Different levels of color quantization are used for generating and evaluating results. For different levels of color quantization the classifier shows improvements in classification accuracy. The average classification accuracy of classifying grade III is 98%. The classification between lower grades has also improved and has reached to the accuracy level of 94 % for lower levels of color quantization.

Table 4.2: Classification accuracies for discriminating histopathological grades using low-level color texture analysis for different levels of color quantization for Test and Training Data II.

TEST AND TRAINING DATA – II				
COOCCURENCE MATRIX SIZE	**GRADE I**	**GRADE II**	**GRADE III**	**OVERALL CLASSIFICATION ACCURACY**
8	77.8 %	88.6 %	88.9 %	99.2 %
16	99.4 %	95.2 %	100 %	100 %
32	94.4 %	72.9 %	100 %	99.4 %
64	100 %	68.6 %	100 %	99.8 %
128	94.4 %	100 %	100 %	100 %
256	66.6 %	59.2 %	100 %	98.2 %
512	99.8%	96.8%	96.2%	92.9%

Table 4.2 shows the classification accuracies using training and testing data II. In this there is also no difficulty in classification of higher grades. The lower grades have good classification results for low level of color quantization.

Table 4.3: Classification accuracies for discriminating histopathological grades using low-level color texture analysis for different levels of color quantization for Test and Training Data III.

TEST AND TRAINING DATA – III				
COOCCURENCE MATRIX SIZE	**GRADE I**	**GRADE II**	**GRADE III**	**OVERALL CLASSIFICATION ACCURACY**
8	77.1 %	60.3 %	87.5 %	84.2 %
16	60.2 %	23.3 %	97.2 %	71.5 %
32	40.9 %	61.1 %	87.5 %	71.5 %
64	57.2 %	56.9 %	75.7 %	77.6 %
128	73.4 %	78.6 %	96.3 %	80.1 %
256	69.3 %	78.5 %	92.5 %	77.5 %
512	97.2%	88.8%	100%	97.0%

Table 4.3 also shows the similar results as for the Table 4.1 and Table 4.2. The higher grades can easily be classified and the lower grades show good results for different levels of quantization.

The cooccurence matrices computed using color quantization of H&E stained images proved to be very successful in classification and discrimination of histopathological grades. The classification results have shown the good separation between the lower grades which is difficult to discriminate using segmentation because of their overlap.

CHAPTER 5

CONCLUSIONS AND FUTUTURE DIRECTIONS

The objective of our research is to design a system that will help the pathologists to efficiently differentiate between the different histologic grades of Follicular Lymphoma, a type of cancer of lymphatic system. The results of the proposed method are better in terms of classification accuracy. Our main purpose is to improve the classification accuracy and reduce the readability discrepancies in reproducing the results.

5.1 Conclusion

A system to help and aid the diagnosticians in grading of H&E stained Follicular Lymphoma images has been proposed. The color quantization is done prior to the computation of cooccurence matrix to have features that will accurately discriminate the grades. In addition to this method, a ratio criterion method has also been proposed for classification between lower & intermediate grades and higher grades. The proposed computer assisted system helps the pathologists to classify the input image as grade I, II or III. We have achieved remarkable classification accuracy in

discriminating lower grades from higher grades. Using the textural features we have also eliminate the discrepancy in classification of lower grades.

The two set of features were computed. The first feature set was computed using the segmentation of input images and take the ratio of amount of two main segmented regions, i.e. cytoplasm and nuclei. The second feature set was calculated using cooccurence matrix computed for color quantized image.

The histopathological grading of Follicular Lymphoma images have improved to a greater extent using proposed color texture analysis and quantitative methods. This proposed system has also removed the reader variability and provides consistent results. The overall accuracy of the grades I, II and III is 81.2%, 85.3% and 90.7%, respectively. We have achieved the average classification accuracy of 85.7 %.

5.2 Future Work

The Microscopic Image Analysis is an open field to work further and improvements can be brought in to efficiently detect different types of lesions. The proposed work is designed, implemented and evaluated only for H&E stained Follicular Lymphoma images. For example the proposed method can also be implemented on neuroblastoma cancer and prostrate cancer. Furthermore, the classification accuracy can be enhanced using different combinations of feature sets along with GLCM features computed in the proposed algorithm.

In our future work we will design a system to segment and detect the desired cell components individually and separates them according to their shape, size, area and distance. For this purpose we will use a grid based layout models and semantic graph based methods to extract features useful for discrimination. We will upgrade our proposed system to a three stage system do that the grades can be accurately differentiated.

BIBLIOGRAPHY

[1] G. E. Metter, B. N. Nathwani, J. S. Burke, C. C. Winberg, R. B. Mann, and M. Barcos., "Morphological sub classification of follicular lymphoma: Variability of diagnoses among hematopathologists, a collaborative study between the repository centre and pathology panel for lymphoma clinical studies," *Journal of Clinical Oncology*, vol. 3, pp. 25–38, 1985.

[2] F. Dick, S. Van Lier, P. Banks, G. Frizzera, G. Witrak, and R. Gibson, "Use of the working formulation for non-Hodgkin's lymphoma in epidemiological studies: Agreement between reported diagnoses and a panel of experienced pathologists," *Journal of National Cancer Institute*, vol. 78, pp. 1137–1144, 1987.

[3] The Non-Hodgkin Lymphoma Classification Project. A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin lymphoma. *Blood, 89*, pp. 3909–3918, 1997.

[4] E. S. Jaffe, N. L. Harris, H. Stein, and J. W. Vardiman, Tumours of haematopoietic and lymphoid tissues, *Lyon: IRAC Press,*2001.

[5] G. A. Meijer, J. A. M. Belien, P. J. Diest, and J. P. A. Baak, "Image analysis in clinical pathology," *Journal of Clinical Pathology*, vol. 50, pp. 365–370, 1997.

[6] C. Demir, and B. Yener, "Automated cancer diagnosis based on histopathological images: a systematic survey," *Technical report TR-05-*

09, Rensselaer Polytechnic Institute, Department of Computer Science, Troy, NY1, 2005.

[7] O. Sertel, U. V. Catalyurek, J. H. Saltz, M. N. Gurcan, and G Lozanski, "Computerized microscopic image analysis of follicular lymphoma", *Proc. of SPIE Medical Imaging 2008: Computer-Aided Diagnosis*, 2008.

[8] M. Zorman, P. Kokol, M. Lenic, J. L. Sanchez de la Rosa, J.F. Sigut, and S. Alayon, "Symbol-Based Machine Learning Approach for Supervised Segmentation of Follicular Lymphoma Images," *Twentieth IEEE International Symposium on Computer-Based Medical Systems, CBMS*, pp. 115-120, 2007.

[9] M. N. Gurcan, T. Pan, H. Shimada, and J. Saltz, "Image Analysis for Neuroblastoma Classification: Segmentation of Cell Nuclei," *Engineering in Medicine and Biology Society, 28th Annual International Conference of the IEEE*, pp. 4844-4847, 2006.

[10] G. Paschos, "Perceptually uniform color spaces for color texture analysis: an empirical evaluation," *IEEE Transactions on Image Processing*, vol. 10, no. 6, pp. 932 – 937, 2001.

[11] A. Robertson, "The CIE 1976 Color-Difference Formula," *Color Research and Application*, vol. 2, no. 1, pp. 7-11, 1977.

[12] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (2nd Edition)*. Prentice Hall, January 2002.

[13] J. P. Braquelaire, and L. Brun, "Comparison and optimization of methods of color image quantization," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 1048–1052, 1997.

[14] Anthony H. Dekker, "Optimal color quantization using Kohonen neural networks," *Technical Report, TR 10/93, Department of Information System and Computer Science, National University of Singapore*, October 1993.

[15] C. H. Chang, P. Xu, R. Xiao, and T. Srikanthan, "New adaptive color quantization method based on self-organizing maps," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 237 – 249, 2005.

[16] S-C Pei, and Y-S Lo, "Color image compression and limited display using self-organization Kohonen map" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 2, pp. 191-205, 1998.

[17] Y. Wu, C. Yang, and T. Wang, "A new approach of color quantization of image based on neural network," *International Joint Conference on Neural Networks, Proceedings. IJCNN '01*, vol. 2, pp. 973-977. 2001.

[18] T. Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464 – 1480, Sep 1990.

[19] T. Kohonen, *Self-organizing maps*, Heidelberg: Springer, 1997.

[20] R. W. Conners, and C. A. Harlow, "A Theoretical Comparison of Texture Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 3, pp. 204-222, 1980.

[21] R. M. Haralick, "On a texture-context feature extraction algorithm for remotely sensed imagery," *IEEE Conference on Decision and Control,* vol. 10, pp. 650-657, 1971.

[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 3, no. 6, pp. 610-621, 1973.

[23] I. Ashdown, "Octree Color Quantization," *from the book: Radiosity-A Programmer's Perspective*, Wiley, New York, 1994.

[24] P. Heckbert, "Color image quantization for frame buffer display," *Computer & Graphics*, vol. 3, pp. 297-307, 1982.

[25] N. Papamarkos, A. Atsalakis, and C. Strouthopoulos, "Adaptive Color Reduction," *IEEE Trans. On Systems, Man, and Cybernetics,* Part B, vol. 32, pp. 44-56, 2002.

[26] J. M. Buhmann, D. W. Fellner, M. Held, J. Kettere, and J. Puzicha, "Dithered Color Quantization," *Proceedings of the EUROGR.'98 Computer Graphics Forum*, vol. 17, no. 3, pp. 219-231, 1998.

[27] N. Papamarkos, "Color reduction using local features and a SOFM neural network," *Int. Journal of Imaging Systems and Technology*, vol. 10, pp. 404-409, 1999.

[28] V. Arvis, C. Debain, M. Berducat, and A. Benassi, "Generalization of the co-occurrence matrix for color images: Application to color texture classification," *Image Analysis and Stereology,* vol. 23, no. 1, pp. 63–72, 2004.

[29] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* John Wiley & Sons Inc.

[30] P. A. Devijver, and J. Kittler, *Pattern Recognition: A Statistical Approach.* Prentice Hall, Englewood. Cliffs, NJ, 1982.

[31] O. Sertel, J. Kong, U. V. Catalyurek, J. H. Saltz, M. N. Gurcan, and G Lozanski, "Histopathological Image Analysis Using Model Based Intermediate Representations and Color Texture: Follicular Lymphoma", *Journal of Signal Processing Systems,* vol. 55, pp. 169-183, 2009.