2013

Uterine Cervical Cancer Histology Image Feature Extraction and Classification

Cheng Lu

UTERINE CERVICAL CANCER HISTOLOGY IMAGE FEATURE EXTRACTION AND CLASSIFICATION

by

CHENG LU

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

2013

Approved by

R. Joe Stanley, Advisor
Randy H. Moss
William V. Stoecker

ABSTRACT

Cervical cancer, the second most common cancer affecting women worldwide and the most common in developing countries can be cured in almost all patients, if detected early and treated. However, cervical cancer incidence and mortality remain high in resource-poor regions, where early detection systems often cannot be maintained because of inherent complexity. The National Cancer Institute (NCI) has collected a vast amount of visual information, 100,000 cervigrams (35 mm color slides), screening thousands of women by this technique. In addition to the cervigrams, large digitized histology images are being archived.

In this research, a framework for automatic recognition and classification of cervical intraepithelial neoplasia has been developed. Data sets of 62 image sets with segmented squamous epithelium regions were obtained from the National Library of Medicine, which were analyzed using the framework developed.

This thesis presents methods used in this research to improve the classification results by implementing different feature extraction algorithm and classification algorithm. A leave-one-image-out approach was explored and yielded an overall classification rate as high as 72.58% for exact classification scoring using the cervical intraepithelial neoplasia (CIN) classes Normal, CIN1, CIN2, and CIN3.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere appreciation to my advisor, Dr. R. Joe Stanley, for his guidance, inspiration, and for the financial support he has provided throughout the course of my research and studies. Dr. Stanley is the first advisor to introduce me to the Digital Image Processing field, who always teaches me using multiple methods to solve one problem, not just for research related, but also problems in real life. I also wish to thank my committee members Dr. Randy H. Moss and Dr. William V. Stoecker for their guidance and support.

I would like to thank all my friends for sharing all the joyful moments with me. I would like to thank BeiBei Cheng for introducing me to Dr. Stanley, her advisor. I would like to thank all my colleagues for always sharing ideas with me and working with me side by side. I would like to thank Gwen Siu, my first tutor when I came to the United States, not only for teaching me English and grammar, but also for teaching me knowledge in life.

Finally and most importantly, I am extremely grateful to my family members, who have always wished the best for me and showered all their love and affection on me in spite of their physical absence. I would like to thank Uncle Nathan and Aunt Xu for their guidance and support. I would like to thank my father for his encouragement and support. The deepest thanks are expressed to my mother and my step father, who have always been there for me through all my joy and sadness.

This research was supported by National Library of Medicine (NLM).

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

Annually, there are 400,000 new cases of invasive cervical cancer; 15,000 occur in the U.S. alone. Cervical cancer, the second most common cancer affecting women worldwide and the most common in developing countries, can be cured in almost all patients, if detected by high quality repeat Pap screening, and treated. However, cervical cancer incidence and mortality remain high in resource-poor regions, where high-quality Pap screening programs often cannot be maintained because of inherent complexity. An alternative cervical cancer screening uses analysis of visual testing based on color change of cervix tissues when exposed to acetic acid; cervicography is a technique that augments this visual screening by recording a film image of the acetic acid-treated cervix, and has been widely used over the last few decades.

The National Cancer Institute (NCI) has collected a vast amount of visual information, 100,000 cervigrams (35 mm color slides), screening thousands of women by this technique. In addition to the cervigrams, large digitized histology images are being archived; the size of these images may be an order of magnitude, or more, larger than the cervigrams. The long-term objective of the proposed project is to facilitate the development of a unique Web-based database of digitized cervix images for investigating the role of human papillomavirus (HPV) in the development of cervical cancer and its intraepithelial precursor lesions in women. Automatic recognition and classification of cervical intraepithelial neoplasia (CIN) has the ultimate benefit of improving management and reducing healthcare costs for women with cervical neoplasia, a condition associated with high morbidity and mortality risk worldwide. For this it is necessary to design, implement, and test algorithms for classification of epithelium tissue

of the uterine cervix as Normal, CIN1, CIN2, or CIN3, in digitized, histology images in which the epithelium tissue has been segmented.

One commonly used feature to determine the CIN grade is the nuclear-cytoplasmic ratio: a larger ratio corresponds to a more severe CIN degree. For example, atypical cells are seen mostly in the lower third of the epithelium for CIN 1, lower half or two thirds of the epithelium for CIN 2, and full thickness of the epithelium for CIN 3. Figure 1.1 below compares H & E stained examples of the normal, CIN 1, 2, and 3 histology images. Segments of epithelium may be long and may have varying levels of pathology in sub-regions.

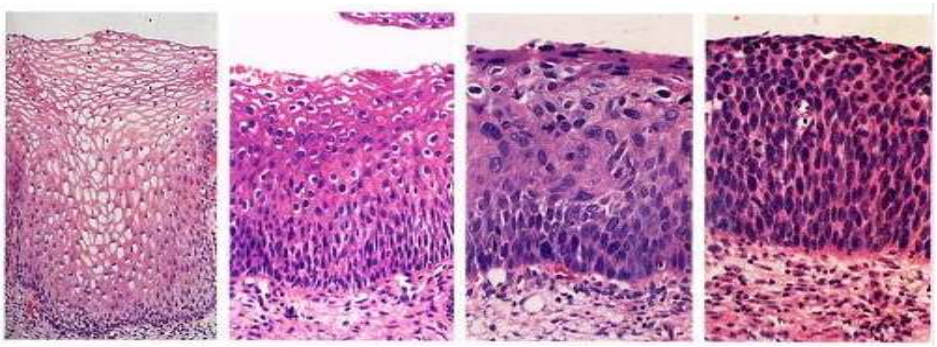


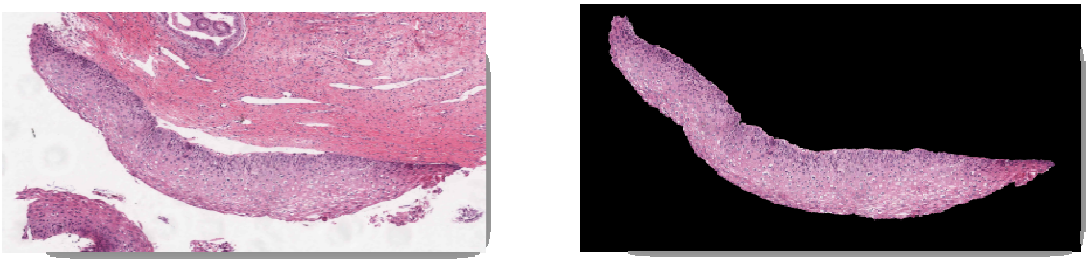
Figure 1.1. CIN examples [1]. (a) Normal; (b) CIN 1; (c) CIN 2; (d) CIN 3.

The remainder of this thesis is organized as follows. Section 2 presents the methodology as a whole for this research. Section 3 focuses on the algorithm for medial axis detection. Section 4 presents a detailed explanation of horizontal and vertical segmentation algorithm and methodology. Section 5 presents the detailed explanation of

every group features. Section 6 presents the classification methodology. Section 7 presents the experiment results and analysis. Section 8 presents the conclusion of this thesis and suggests future scope.

2. METHODOLOGY

The overall goal of the research is to segment the squamous epithelium region from histology slides and classify the squamous epithelial region into different grades of cervical cancer. In this particular study, the squamous epithelium region has been manually segmented by expert pathologists in collaboration with the National Library of Medicine so that these segmented regions could be used as training images for the classification study. Figure 2.1 shows a sample image and its pathologist-segmented epithelial region.



(a) Original image (b) Pathologist segmented image

Figure 2.1. Original image and pathologist segmented image.

However, before going further, it is important to understand what image features can help us classify the images of the squamous epithelium into the four different cervical cancer grades: Normal, CIN1, CIN2 and CIN3. From [1], the most important component that can be used to classify between the different grades of Cervical Intraepithelial

Neoplasia (CIN) is to consider the image to be consisting of three equally spaced segments (from top to bottom) and then analyze each one-third region. In each of these one-third regions, various features can be investigated and then compared across the different classes and the different layers (top third, middle third, bottom third). Features could include texture features, spatial features, color features, etc.

In this study, the method of classifying the images was based on the following steps:

- Medial axis detection, find the medial axis of the segmented epithelium region;
- Image segmentation, divide the segmented image into 10 different vertical blocks along the medial axis, divide the segmented image into 3 horizontal blocks parallel to the medial axis;
- Feature extraction, extract features from each of the blocks;
- Image Classification, classify each of these segmented blocks into the different CIN cervical cancer grades.

The following sections of the thesis will elaborate most of these steps accomplished by author outlined above in detail.

3. MEDIAL AXIS DETECTION

From the manually segmented epithelium regions provided by the National Library of Medicine (NLM), a distance transform-based approach was used for medial axis determination. This technique was developed and implemented by Soumya De [2]. The Matlab 'BWDIST' function was used for computing the distance transform of the segmented region. 'BWDIST' computes the Euclidean distance transform of a binary image. For each pixel in in the binary image, the distance transform assigns a number that is the distance between that pixel and the nearest nonzero pixel. During our experiments, it was found that if the image is rotated first and then the distance transform is applied, the resulting image provides a better representation of the skeleton (used for computing the medial axis) of the binary image. Epithelium region of interest rotation was performed based on estimating the orientation of the object within the image (orientation property within the 'REGIONPROPS' operation in Matlab) and rotating the object with the negative of the orientation angle. The orientation is important because the object's aspect ratio impacts the application of the distance transform, at least for the implementation of the distance transform investigated. A sample of the result of rotating the binary image is shown in Figure 3.1 while the result of applying the 'BWDIST' function on the rotated image is provided in Figure 3.2.

If one can track the line along the brightest pixels of Figure 3.2, this line represents the medial axis of the segmented region. Problems with the end portions of the medial axis using the distance transform makes skeletonization unreliable as a singular medial detection algorithm based on distance transform. The resulting medial axis

obtained from Figure 3.2 is shown in Figure 3.3. This is obtained by taking the maximum values of the pixels (essentially the brightest pixel) along the horizontal axis of the image.

Figure 3.1. Rotating the binary segmented image, (a) Original binary image, (b) Rotated binary image.

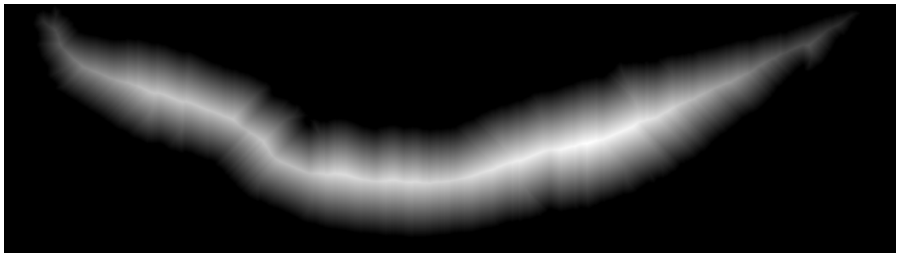


Figure 3.2. Result of applying the distance transform on Figure 3.1(b).

However, a problem was encountered while detecting the medial axis with images that had a somewhat rectangular shape. A few examples of improper detection of the medial axis are shown in Figure 3.4 below. The line shown in pink color is the detected

medial axis using the distance transform approach while the line shown in green is the manually marked medial axis, which is the desirable medial axis.

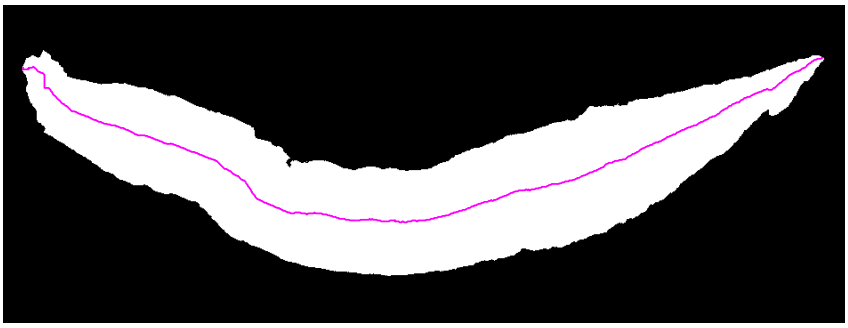


Figure 3.3. Medial axis of the image obtained from Figure 3.2.

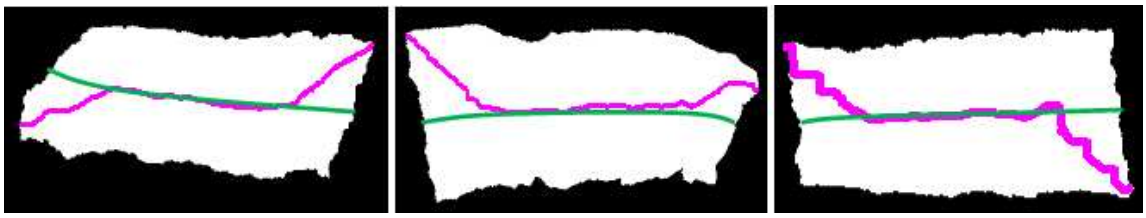


Figure 3.4. Examples of improper medial axis detection using the distance transform approach.

The solution to the above problem is solved by the bounding box-based method. The bounding box-based method is mainly based on ratio comparison of the number of nuclei distributed over 8 masks that are created from the bounding box and control points marked on it. Also for precision purposes a 16-mask approach along with the symmetry factor of the image were taken into consideration. However reports from that process are

pending and would be included in future reports. Figure 3.5 explains the concept. Figure 3.6 shows an example of the medial axis found using the bounding box-based medial axis estimation algorithm.

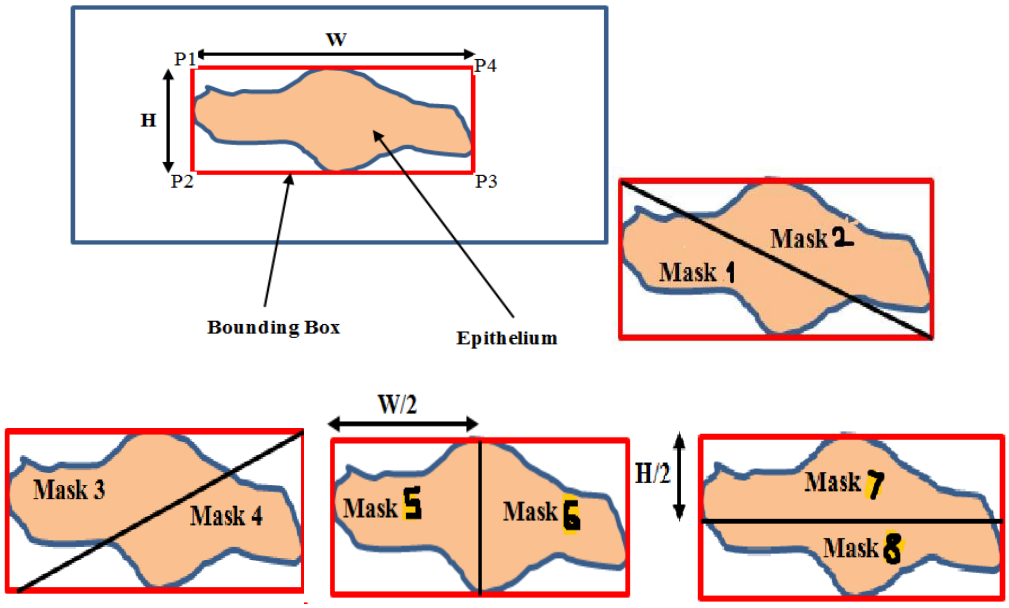


Figure 3.5. The bounding box-based method.

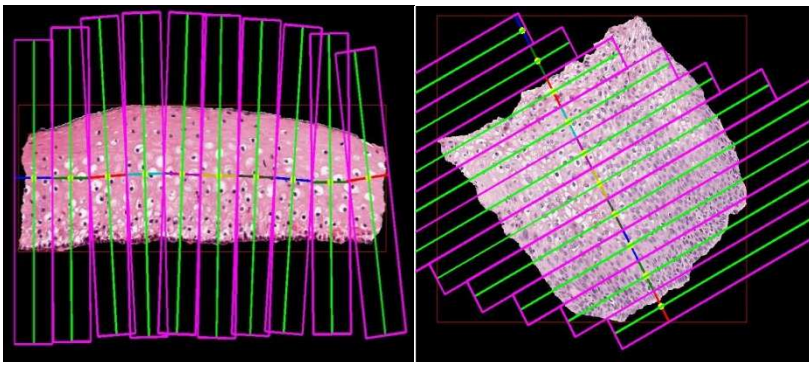


Figure 3.6. Examples of medial axis found using bounding box-based medial axis estimation algorithm.

4. IMAGE SEGMENTATION

There are two reasons that image segmentation is considered a critical step for this research. First, among the 62-image data set, most of the images contain a large number of data points, which gives a lot of burden on the future step, feature extraction. Second, by visually examining the CIN grades, one portion of an image might be considered as normal, while the other portion of the same image might be considered cancerous. Among normal, CIN 1, CIN 2, and CIN 3 grades, one single image might contain multiple grades, which gives inconsistent results for the classification process. By segmenting the images into small segments, more testing images will be collected for further research and each segmented image gives a fairly consistent CIN grade.

4.1. VERTICAL IMAGE SEGMENTATION

Figure 4.1 shows bounding box regions based on breaking the medial axis within the epithelial region into 10 vertical segments. This technique was developed and implemented by Soumya De. The orientation of each segment is determined by taking all of the medial axis points and estimating the slope of the medial axis within each of these ten segments. This is done by first dividing the medial axis points into ten segments. Next, the points within each of the segments are curve-fitted using the 'POLYFIT' function in Matlab. This function uses a least-squares approach to fit the points along the medial axis line. The order of the function was set to 1, meaning that the medial axis points were least-square fitted to a straight line. The corresponding perpendicular orientation is determined so that a bounding box can be generated. For each epithelial region, the 10 bounding box areas (regions) are extracted to be used for feature and classification analysis. The various steps of this method are shown in Figure 4.2 below.

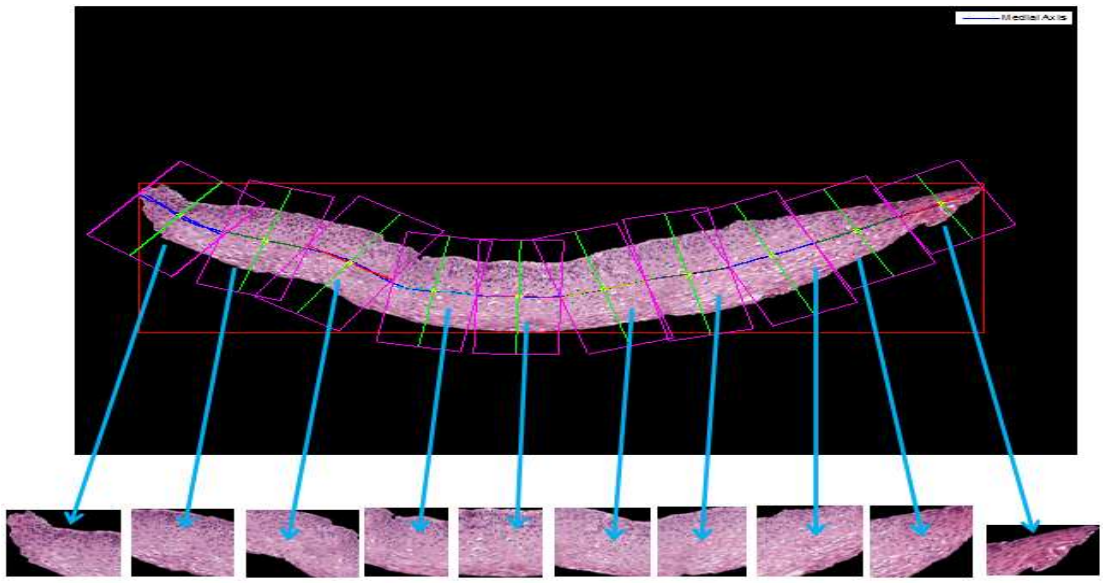
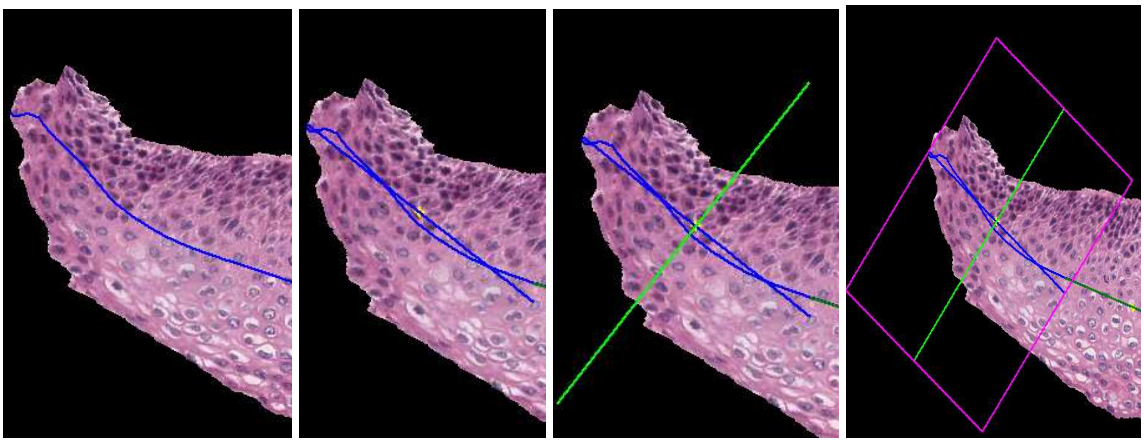


Figure 4.1. Example of medial axis broken into 10 segments with bounding boxes shown/determined for each segment. The boxes are rectangular deviations from 90 degree are due to aspect.



(a)	(b)	(c)	(d)
The portion of the medial axis within the first segment	The least squares fit line obtained from the medial axis points.	The perpendicular of the least squares fit line obtained to generate the bounding box	The bounding box is generated using the information obtained from Steps (a)-(c).

Figure 4.2. The various steps in creating the ten different segments from the epithelium region.

By segmenting every image into 10 segments, a total of 620 segmented images are obtained. The class label is given to individual segmented image based on visually observe the CIN grade.

4.2. HORIZONTAL IMAGE SEGMENTATION

In addition to the vertical segmentation approach, outlined in Section 4.1, an original contribution of this thesis is the development of a horizontal segmentation based method. In this case, the color image was segmented into three parts horizontally, along the medial axis. As shown in Figure 4.3, the first segment consists of the top thirty percent of the epithelium; the second consists of the middle forty percent of the epithelium, while the third segment contains the bottom thirty percent of the epithelium. In order to determine each of these three regions, the following steps were performed.

Step 1: Segment the background and foreground of the original color image, create a binary mask image with background being zeros and foreground being ones.

Step 2: Create a binary image of medial axis line.

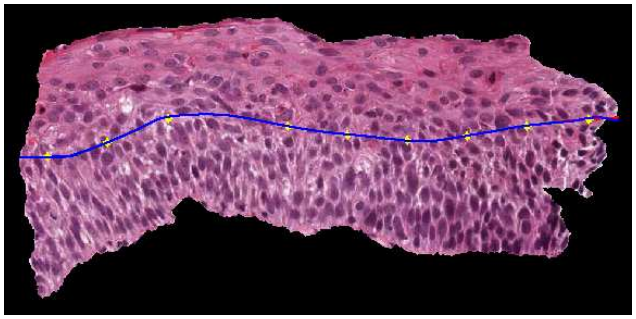
Step 3: Create two binary images, top-half-mask and bottom-half-mask, from the binary mask image by separating the region of interest with medial axis line.

Step 4: Perform dilation using a disk of radius 1 on the medial axis line in the binary image for medial axis line.

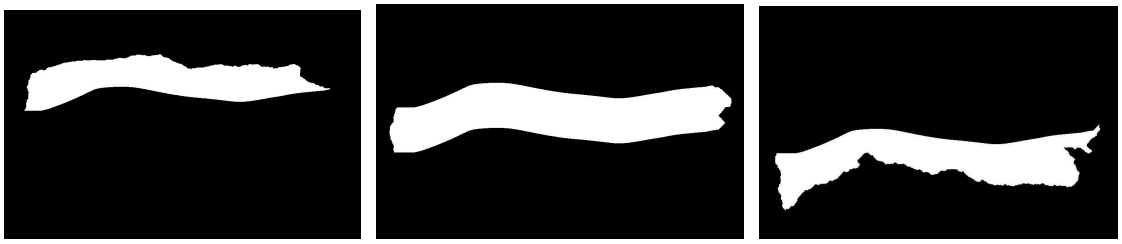
Step 5: Repeat step 4 until the dilated area is 40% of the binary mask image, by counting the total number of pixels those are ones. The middle-third-mask is created.

Step 6: The top-third-mask is created by subtracting the middle-third-mask from top-half-mask. The bottom-third-mask is created by subtracting the middle-third-mask from bottom-half-mask.

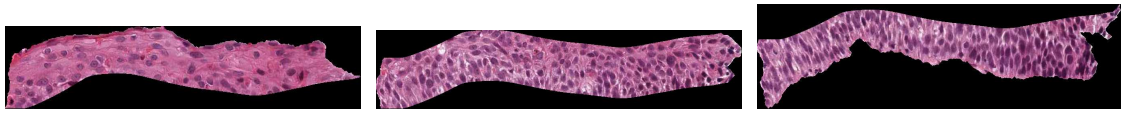
Step 7: After finding the top-third-mask, middle-third-mask, and bottom-third-mask as shown in Figure 4.3 (b), the segmented regions can be found from the original color image as shown in Figure 4.3 (c).



(a) Original image



(b) Segmented masks



(c) Segmented regions

Figure 4.3. Horizontal segmented images.

5. FEATURE EXTRACTION

The feature extraction algorithm is different between vertical segmented image and horizontal segmented image. Every vertical segmented image can still be used as test image, which a class label can be given. On the other hand, every horizontal segmented image cannot be considered as CIN image, which a suitable class label cannot be given. For either case, the feature extraction algorithm extracts seventy-seven features from every segmented image and store them in excel file.

For each vertical or horizontal segment, seven different types of features were investigated. These include: (a) Texture Features, (b) Color Features, (c) Geometry (Triangle) Features, (d) Weighted Density Distribution Features, (e) Nuclei Feature, (f) Light Area Features and (g) Combined Features. Features were extracted from each of the vertical/horizontal segmented regions as obtained from the steps explained in Section 4. Table 5.1 shows the feature numbers and a brief description of each feature while the following sections elaborate the features.

Table 5.1. Feature table.

Feature set	Label	Measure	Description
Texture Features	F1	Contrast of segment	Returns a measure of the intensity contrast between a pixel and its neighbor over the whole segment.
	F2	Energy of segment	Measures the entropy (squared sum of pixel values in the segment)
	F3	Correlation of segment	Returns a measure of how correlated a pixel is to its neighbor over the whole segment.
	F4	Homogeneity of a segment	Returns a value that measures the closeness of the distribution of pixels in the segment to the segment diagonal.
	F5-F6	Contrast of GLCM	Measure of the contrast of the GLCM matrix obtained from the segment.
	F7-F8	Correlation of GLCM	Returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.
	F9-F10	Energy of GLCM	Returns the sum of squared elements in the GLCM.
	F11	Correlation of GLCM	Returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Table 5.1. Feature table (cont.).

Color Features	F12	Percentage Red	Percentage of region that has the reddish pixels.
	F13	Percentage White	Percentage of region that has the whitish pixels.
	F14	Percentage Black	Percentage of region that has the blackish pixels.
Triangle Features	F15	Average area of triangles	This is the average area of the triangles formed by using Delaunay triangulation on the cells detected.
	F16	Std deviation of area of the triangles	This is the standard deviation of the area of the triangles formed by using Delaunay triangulation on the cells detected.
	F17	Average edge length	This is the mean of the length of the edges of the triangles formed.
	F18	Std deviation of edge length	Standard deviation of the length of the edges of the triangles formed.
Correlation-based Features (WDD Features)	F19~F30	Weighted density distribution for whole segment	Correlation of red plane profile of the segment and WDD function for whole segment.
	F31~F42	WDD for top third of the segment	Correlation of red plane profile of the segment and WDD function for top third of the segment.
	F43~F54	WDD for middle third of the segment	Correlation of red plane profile of the segment and WDD function for middle third of the segment.
	F55~F66	WDD for bottom third of the segment	Correlation of red plane profile of the segment and WDD function for bottom third of the segment
Nuclei Features	F67	Average nuclei area	Returns the ratio of total nucleus area over total number of nuclei
	F68	Ratio of background area over nucleus area	Returns the ratio of total background (Nuclei) area over total nucleus area
Light Area Features	F69	Ratio RGB	Returns the average intensity of RGB image over background
	F70	Ratio R	Returns the average intensity of R-plane in luminance image over background
	F71	Ratio G	Returns the average intensity of G-plane in luminance image over background
	F72	Ratio B	Returns the average intensity of B-plane in luminance image over background
	F73	Ratio LUM	Returns the average intensity of L-plane in luminance image over background
	F74	Unit size of light area	Returns the number of light area over total area
	F75	Ratio of light area over background area	Returns the ratio of total light areas over total background (Light) area
Combined Features	F76	Ratio of light area number over nuclei number	Returns the ratio of light area number over nuclei number
	F77	Ratio Light over Nuclei	Returns the ratio of total light areas over total nuclei area

5.1. TEXTURE FEATURES

Eleven texture features were computed for this study for each of the vertical/horizontal segmented blocks [3]. These included the contrast (F1) , energy (F2), correlation (F3) and uniformity (F4) of the segmented region, combined with the same statistics obtained from the gray level co-occurrence matrix (GLCM) formed from the segment (F5-F11, see Table 5.1). Details of the GLCM method can be found in [4].

5.2. COLOR FEATURES

For computing the color features, three different areas were marked as Red region, Black Region and White region within one of the sample images (ouhsc_d26-2-cin3.jpg), which was selected randomly. As shown in Figure 5.1, the Red region is representative of pixels which are reddish in nature while the Black and White regions represent regions which are blackish and whitish in nature, respectively. The average pixel values for these three regions were computed and used as cluster centers clustering the segmented region into three different regions (Red, Black, and White). Figure 5.1 shows an example image and the Red, Black and White clustered regions.

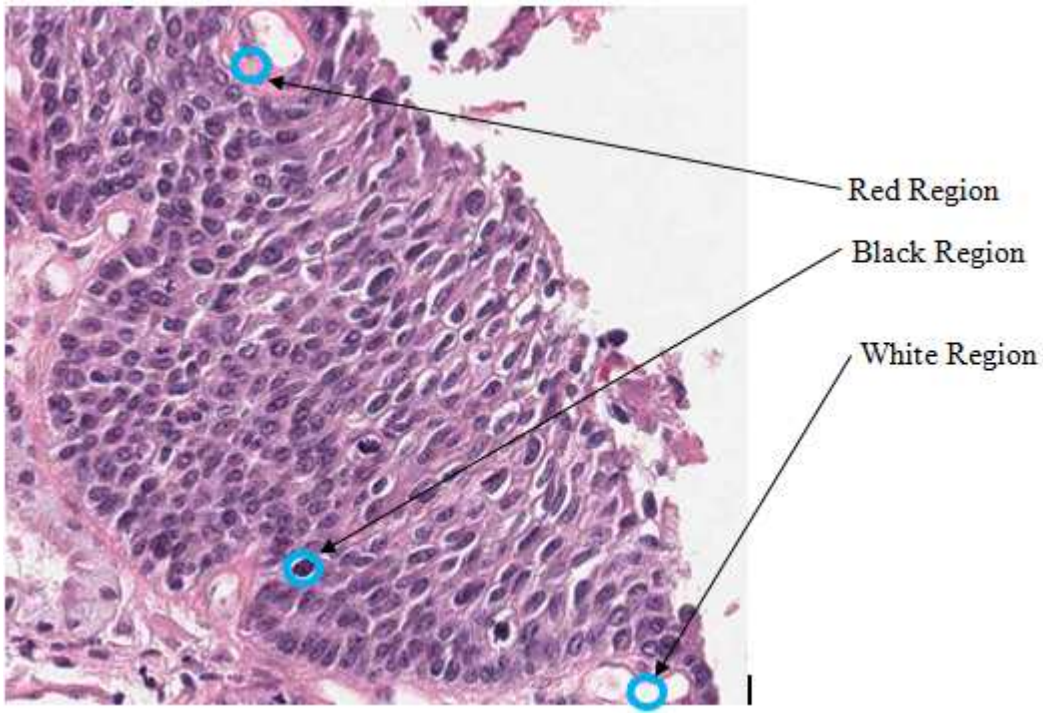


Figure 5.1. Representative color regions within an image.

As shown in Figure 5.2, from the Red, White and Black regions, the color features are calculated as percentage of Red Region (F12), percentage of White region (F13) and percentage of Black region (F14) within the segmented epithelium region.

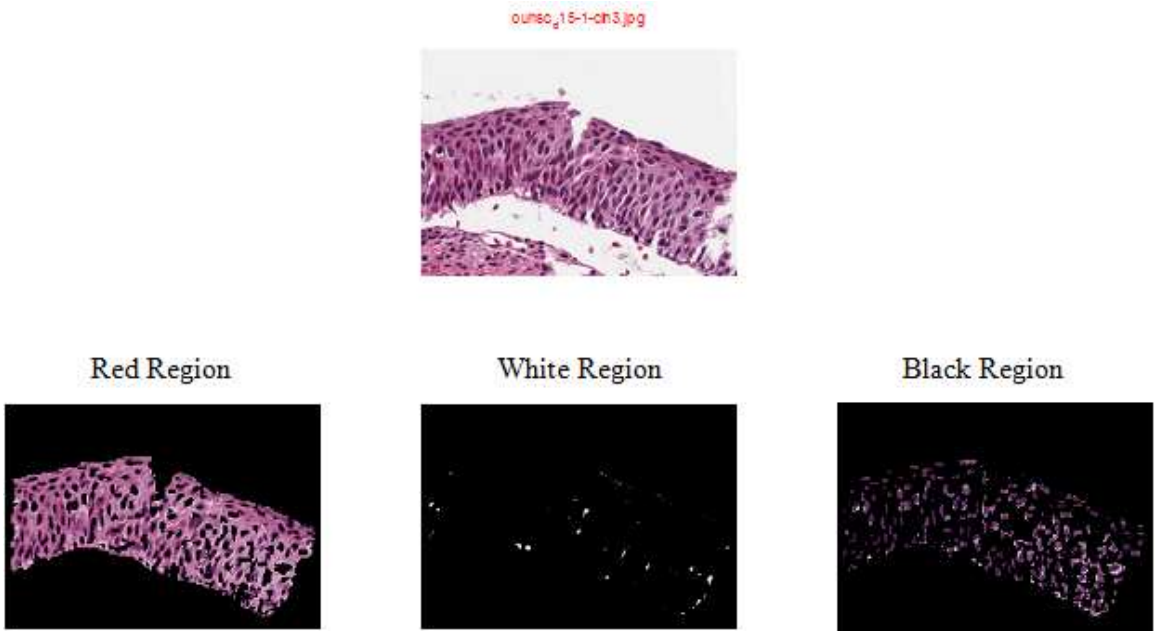


Figure 5.2. Color clusters obtained for color features computation.

5.3. GEOMETRY (TRIANGLE) FEATURES

Previous research has shown that the triangles formed by joining the centroid of the cells detected can provide information on the nature of the squamous epithelium [5]. In the current work, the circular Hough Transform based circle detection was used for detecting the cells. This method is based on the assumption that the smallest structure that defines the progressively increasing cell size is a circle. The idea is to use the circle as a

simplified structure of the cell. Once the cells have been detected using the circular Hough Transform, all the cells are joined together to form triangles using the Delaunay Triangulation method. The Matlab function 'DELAUNAY' was used for the triangulation.

As shown in Figure 5.3, for each segmented region, the cells are located using the circular Hough Transform and then the triangles are generated. The features that are obtained from the triangles include: average area of the triangles (F15), standard deviation of the area of the triangles (F16), average distance between the lengths of edges of the triangles found (F17) and standard deviation of the distance between the lengths of edges of the triangles (F18).

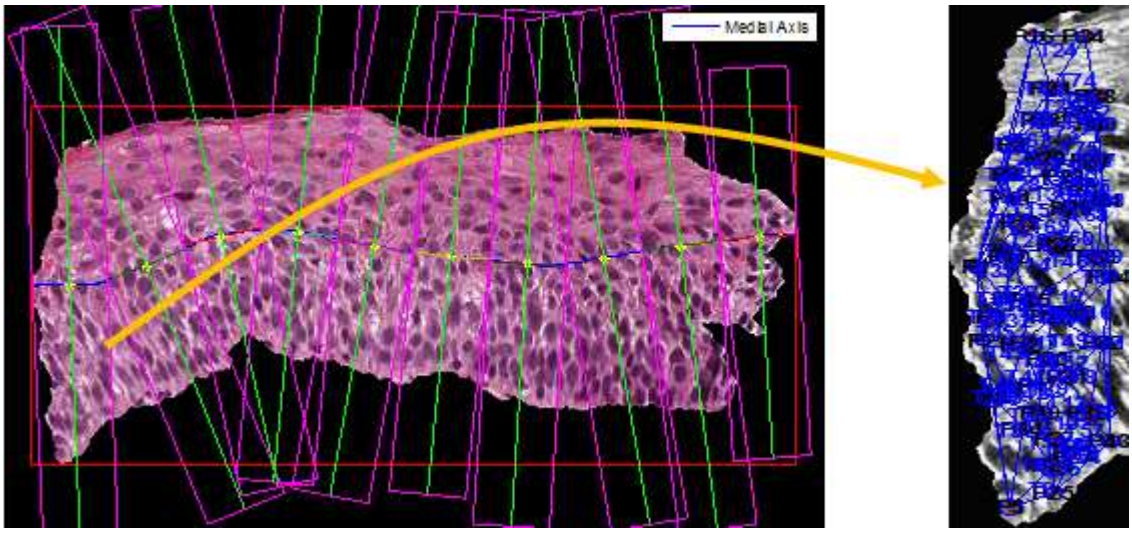


Figure 5.3. Triangles formed from the segments.

5.4. WEIGHTED DENSITY DISTRIBUTION FEATURES

The fourth set of features is based on computing texture profiles of the segments obtained using the medial axis approach and correlating those profiles with basic functions. The texture profile of each segment is found as follows. Let SR and SC denote the rows and columns of the segment, respectively. The profile value for each row, P (i), of the segment S is defined as Equation 3. S (i, j) represents the red plane pixel value at ith column and jth row.

$$\boldsymbol{P(i)} = \frac{\sum_{j=1}^{S_C} \boldsymbol{S(i,j)}}{\boldsymbol{S_C}} \quad \text{for } i = 1,\ldots, S_R. \tag{1}$$

Let $\boldsymbol{P} = \{\boldsymbol{P}(1), \boldsymbol{P}(2), \ldots, \boldsymbol{P}(\boldsymbol{S_R})\}$ be the sequence of profile values. Correlation-based features are extracted by correlating the red plane pixel value profile of the segment with weighted density distribution (WDD) functions [7], shown in Figure 5.4. Let W_1 denote the WDD function in Figure 5.4(a), W_2 denote the WDD function in Figure 5.4(b), and so on.

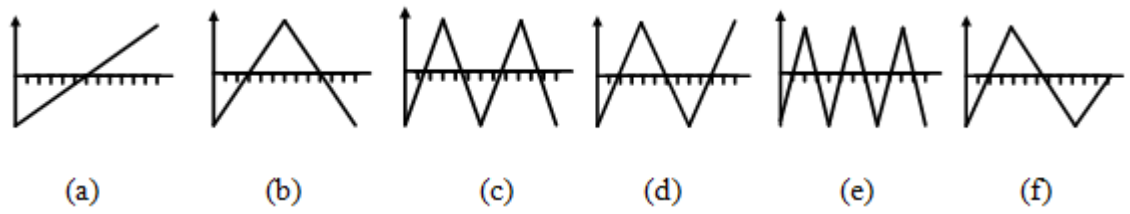


Figure 5.4. The WDD functions used (adapted from [7]).

The twelve correlation-based features are computed as follows:

Six WDD features $(f_1, \ldots, f_6)$ $f_1, f_2, \ldots f_6$ are computed using the profile *P* according to the following expression:

$$f_k = \sum_{i=1}^{S_R} P(i) W_k(i) \quad \text{for } k = 1, 2, \ldots 6. \tag{2}$$

$k = 1,2, \ldots ,6$.Six additional features $f_7, f_8, \ldots f_{12}$, are computed by correlating the six WDD functions with the sequence of absolute differences between samples value as follows:

$$f_{k+6} = \sum_{i=1}^{S_R} \left| P(i) - P(i-1) \right| W_k(i) \quad \text{for } k = 1, 2, \ldots 6 \text{ and } L(0) = 0. \tag{3}$$

A total of 48 WDD features are obtained using the above method for the following four different variations of the segment under analysis: a) whole segment (F19-F30) (from row 1 of the epithelium to row SR), b) top third of the segment (F31-F42) (from row 1 to SR/3), c) middle third of the segment (F43-F54) (from row SR/3 + 1 to 2SR/3), and d) bottom third of the segment (F55-F66) (from row 2SR/3 + 1 to SR). Note the distinction between top, middle, and bottom third here refers to the actual segment or block, which could be one of the 10 vertical segments.

5.5. NUCLEI FEATURES

For the rest of this section, The paper emphasis on the recent feature development. This part of the feature extraction was developed and implemented by

Cheng Lu and Peng Guo [8]. After many trials and errors, a nuclei detection algorithm is tested based on epithelium image pre-processing. The goal is to enhance the image before feature extraction, which allow the nuclei detection algorithm detect nuclei much easier. These pre-processing procedures include averaging, image sharpening, histogram-equalization, high boosting, etc. There are two steps in this segmentation process, image enhancement and nuclei detection.

5.5.1. Nuclei Feature Pre-processing. Before nuclei detection, we take a step of image enhancement of gray scale image. There are many different approaches for image enhancement. For this project, a variety of filters are applied to the images, including Laplacian, Canny, Roberts and Sobel as shown in Figure 5.5.

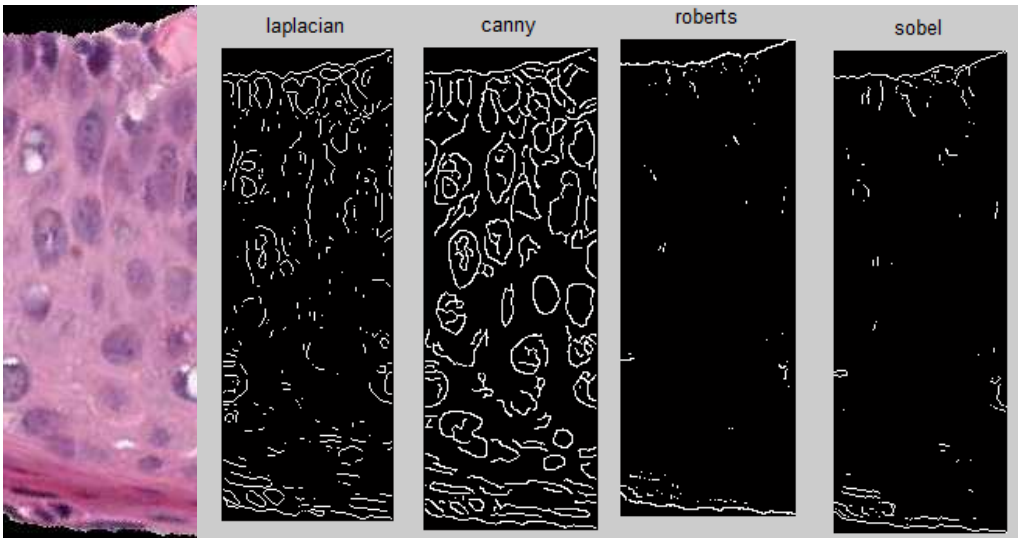


Figure 5.5. Original image and edge detector images.

An image enhancement process called high boost filtering [9] is used to improve the contrast between the nuclei and the background as shown in Figure 5.6. The high boost filter made the edges of the nuclei more distinguishable.

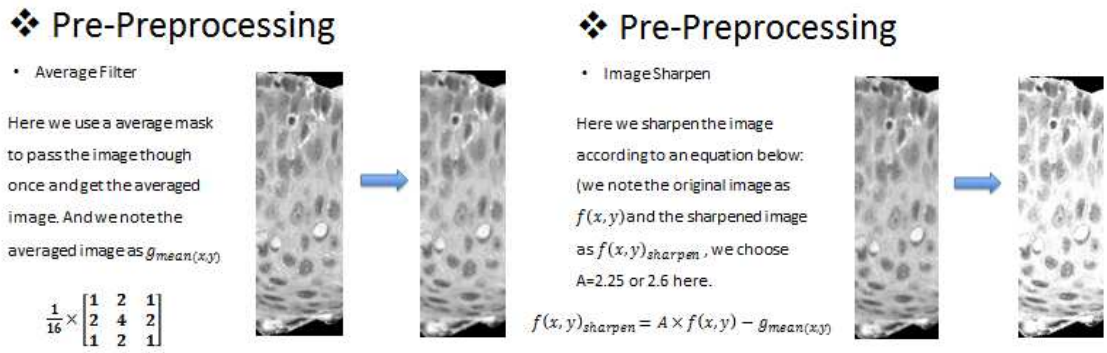


Figure 5.6. Method of image sharpening.

$$I_{sharpen}(x,y) = A\,f(x,y) - g(x,y) \tag{4}$$

From equation 4 above, the sharpened image $I_{sharpen}$ sharpen at pixel location (x, y), f (x, y) is the original image in gray scale. g (x, y) is calculated from passing an averaging filter to the gray scale image. A is varied by the result from the output image. A was determined to be 2.25 based on analysis of experimental data set. The I_sharpen (x, y) is the resulting image as shown in Figure 5.6. After using High boost filter, histogram equalization is applied to the image. A histogram equalization algorithm is used as the next step for image enhancement. Figure 5.7 shows the image before and after histogram equalization process. This step equalizes the values in each pixel, so the range of values equally distributed from 0 to 255 as shown in Figure 5.8. Before using histogram equalization, the distributions of pixel numbers are not equally distributed.

Most of the pixel values are in the range of 100 to 200, or the gray area. After using histogram equalization, the distributions of pixel numbers are equally distributed. Figure 5.8 shows the histogram that every pixel value has approximately equal amount of pixel numbers.

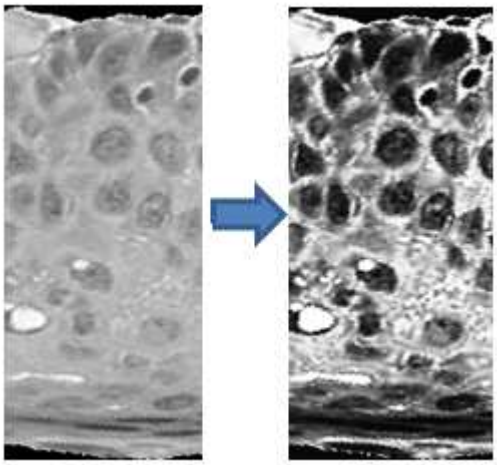


Figure 5.7. Applying histogram equalization (before and after).

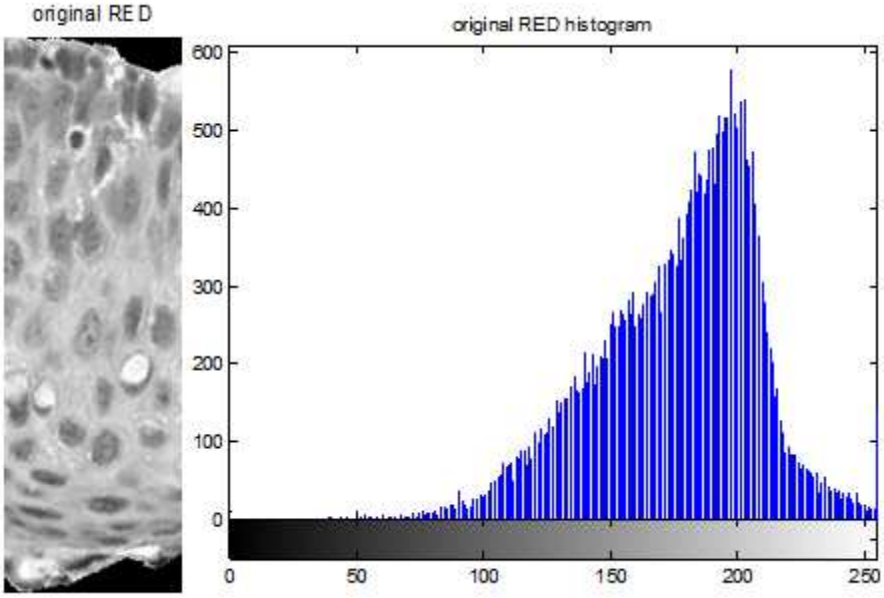


Figure 5.8. Applying histogram equalization (before and after).

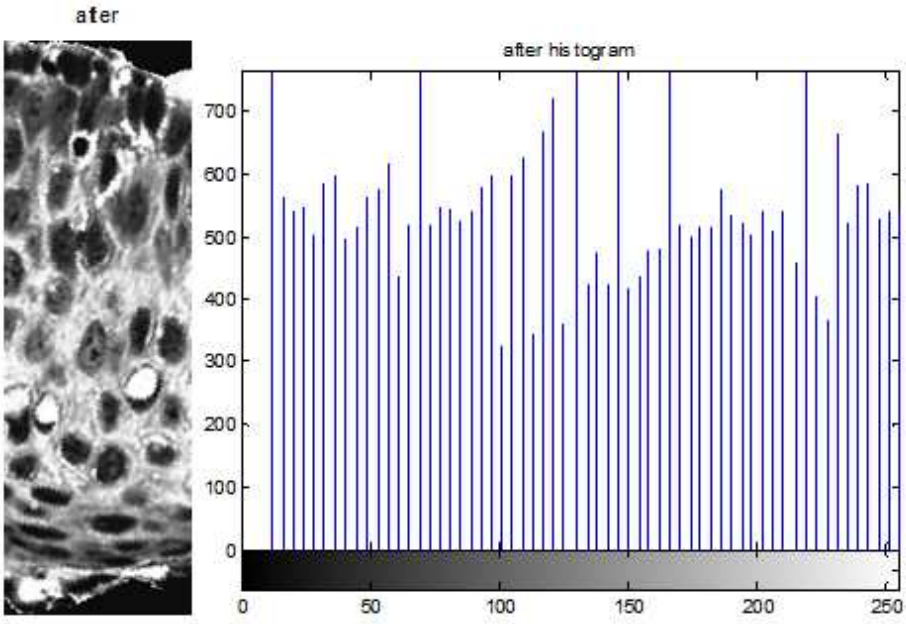


Figure 5.8. Applying histogram equalization (before and after) (cont.).

5.5.2 Nuclei Region Segmentation (Nuclei Processing). After testing and correction, a portion of the nuclei detection code supplied by NLM is used to perform nuclei processing. The algorithm has many progresses such as clustering, holes filling, small-area eliminating [9], etc., which is shown in Figure 5.9.

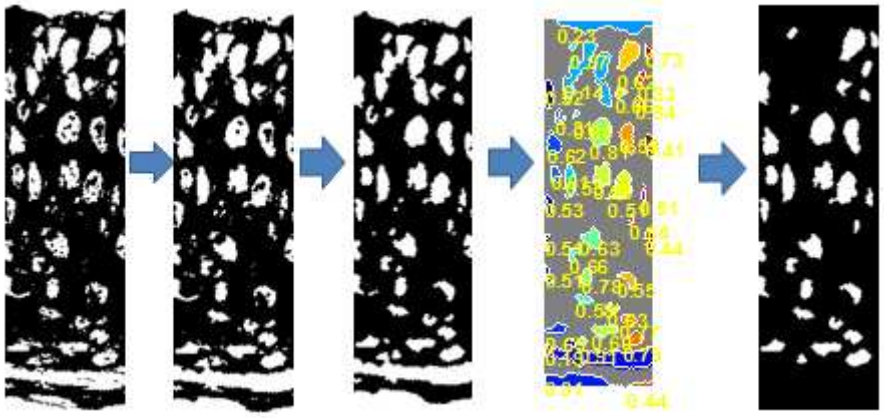


Figure 5.9. Nuclei detection Progress (a) - (e).

The clustering process is shown in Figure 5.9 (a) from histogram equalization result. A region set of points which have the similar characteristics was grouped into one cluster. Figure 5.9 (b) shows the result after holes filling process with Matlab function imclose. Figure 5.9 (c) shows the result after small-area eliminating process with Matlab function imopen. Figure 5.9 (e) shows the result of nuclei detection algorithm. The algorithm keeps objects those have rounded shapes, or compactness. The compactness describes how compacted the cluster is. Compactness was determined to be Equation 5.

$$Compactness = \frac{4 \times \pi \times Area}{(Perimeter)^2} \tag{5}$$

The algorithm keeps the clusters with compactness more than 0.5 and eliminates the cluster with compactness less than 0.5 based on analysis of experimental data set.

Since the contrast of the images is improved, the nuclei detection code can produce a better result. The initial plan for this project is only processing the red layer of the RGB image. After examination, the green and blue layers give similar but slight different resulting images as shown in Figure 5.10. Even though three layers have very similar result, but none of them gives a conclusive result. It is possible to have a better result by combining all three layers, but the processing time is the only disadvantage for this approach. It takes three times more calculation time to process three layers in comparison one. After empirical analysis of the color image data set, only the red color plane was examined. The features that are obtained from the Nuclei include: average nuclei areas, the ratio of total nucleus areas over total number of nuclei (F67), ratio of background area over nucleus area (F68).

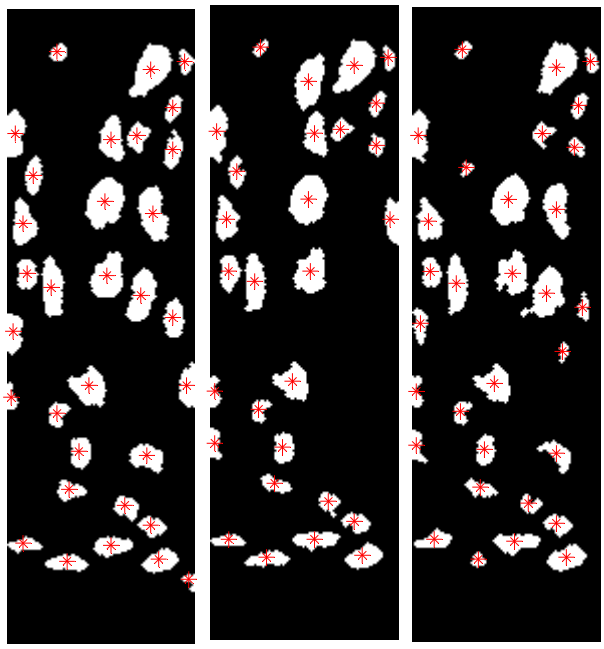
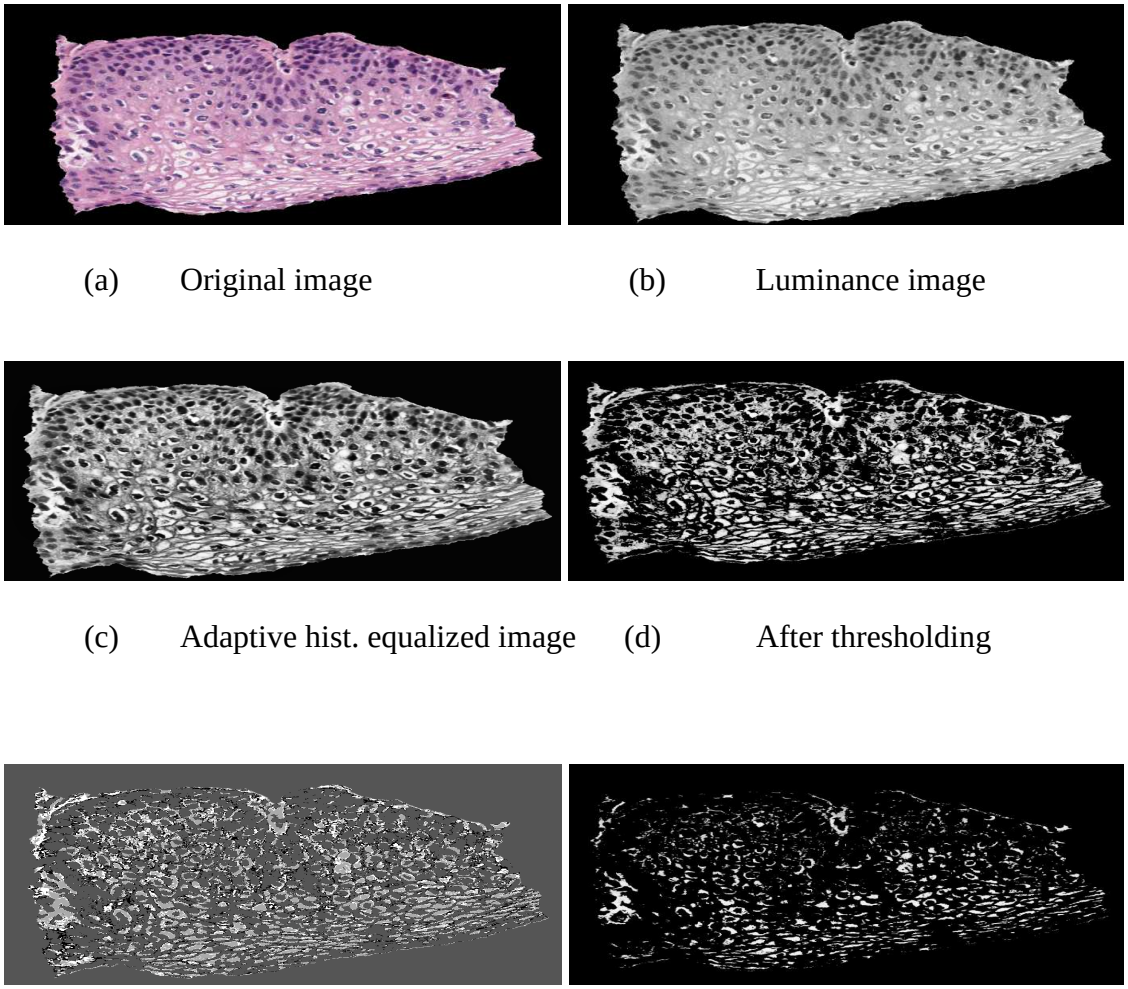


Figure 5.10. Segmented nuclei, (a) Red layer, (b) Green layer, (c) Blue layer.

5.6. LIGHT AREA FEATURES

This part of the feature extraction is developed and implemented by Koyel Banerjee and Xiao Pan [8]. The challenge that goes with extracting the light area regions from the original image is mainly the color and intensity variations. Often the light areas are mistaken for white areas which are not the case. The light areas may appear white to the human eye, but the light areas tend to be more on the tail on the histogram where there is the concentration of light areas or high intensity values. Also the other problem faced was that the light areas do not have a pre-defined shape like the nuclei so we cannot take into account the shape/morphology of these regions. Therefore, to avoid such shortcomings an attempt to process these regions in the color plane was done taking into account the a-plane and b-plane and discarding the L-plane. The L-plane provided the best visual results of the 3 planes examined.

It is to be noted that the light-area concentration in the normal image is much more than its concentration in CIN1 or CIN2 or CIN3. So in general high light-areas are found in Normal and CIN1 type images and it is very less in CIN2 types and barely noticeable in CIN3 types of images. The step-by-step results are shown in Figure 5.11.



(a) Original image (b) Luminance image

(c) Adaptive hist. equalized image (d) After thresholding

(e) Image labeled by cluster_idx (f) Final cluster

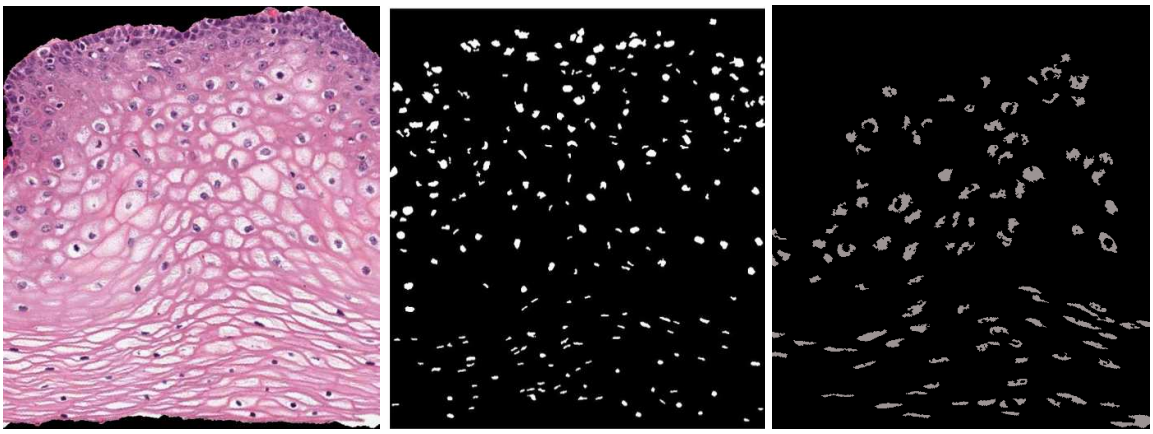
Figure 5.11. Light-area segmentation process.

(g) Final light areas for a cin1 image after dilation & erosion

Figure 5.11. Light-area segmentation process (cont.).

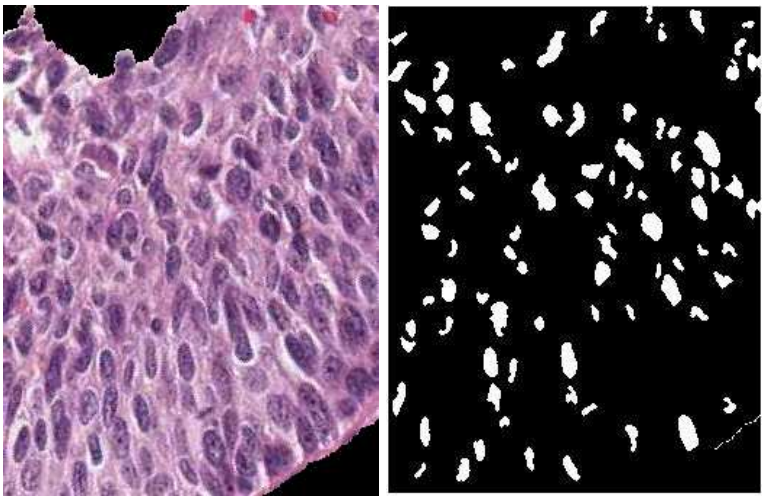
5.7. COMBINED FEATURES

Using the algorithms for segmenting nuclei and light areas, new features have been developed for vertical segment and image-based classification. The algorithm successfully extracts the number of nucleus, the total area of nucleus, the number of light areas and the total light area. With additional calculation, different ratio between nucleus and light areas are found. Each feature mentioned in above is presented in Table 5.1. In comparison as shown in Figure 5.12, a normal image has fewer nuclei than CIN 3 image and a normal image has more light area than CIN 3 image.



(a) Normal (b) Nuclei Mask (c) Light-Area Mask

Figure 5.12. Comparison between nuclei mask and light-area mask.



(d) CIN 3 (e) Nuclei Mask (f) Light-Area Mask

Figure 5.12. Comparison between nuclei mask and light-area mask. (cont.).

6. CLASSIFICATION

This thesis covers different stages of a continuation research. New approaches related to any of the previous sections are constantly introduced and analyzed. If the old approaches consistently produce unsatisfied result, ones will be stop experimenting in the future research. Since image segmentation involves both horizontal segmentation and vertical segmentation, the features extracted from them need to be stored and processed separately. The vertical segment (10 segments) features involves a total of 620 images with 77 features, which is stored in a 620 by 77 matrix in Excel format. Horizontal segmentation, as described in Section 4.2, uses a different approach. Since every image is going to be segmented into top third, middle third and bottom third, each segmented image cannot be used as a standalone image. The features computed from each of the horizontal segments into a single feature vector for image-based classification. In the early stage of the research, NLM supplies 29 images, while the next 33 images are supplied in the later research, providing a total of 62 images examined in this research.

The following approach was used for image-based classification into one of the CIN classes (Normal, CIN1, CIN2, and CIN3).

For vertical segmented image classification, we carried out the following four steps:

Step 1: Train the segments with different classifiers, support vector machine (SVM), linear vector quantization (LVQ), and Bayes classifier using a leave-one-image-out approach. The classifier is trained based on the individual vertical segment feature vectors for all but the left-out epithelium image (used as the test image).

Step 2: Classify each vertical segment of the left-out test image into one of the CIN grades using the linear discriminant analysis (LDA classifier).

Step 3: Assign the test epithelium image to the class (Normal, CIN1, CIN2, CIN3) using a voting scheme. The CIN grade of the test epithelium image was assigned based on whichever class is most frequently assigned to each of the vertical segments for the image (most frequently occurring class assignment for the ten vertical segments in Step 2). If there is a tie with the most frequently occurring class assignment among the vertical segments, then the epithelium image is assigned to the higher class. For example, if there is a tie between CIN2 and CIN3, then the image would be labeled as CIN3.

Step 4: Repeat steps 1-3 for all the epithelium images in the experimental data set.

The horizontal features data involves 29 images with 90 features (Feature 1 to 30 from each horizontal segmented image), which is stored in a 29 by 90 matrix in Excel format.

The following approach was explored for image-based classification using the horizontal segments:

Step 1: Train the classification algorithm (SVM, LVQ, Bayes) using a leave-one-image-out approach. The classifier is trained based on the 28 individual horizontal segment feature vectors for all but the left-out epithelium image (used as the test image).

Step 2: Classify the left-out test image into one of the CIN grades using the different classification algorithm (SVM, LVQ, and Bayes).

Step 3: Repeat steps 1-2 for all the epithelium images in the experimental data set.

A detailed analysis compare between three classifiers is shown in Section 7. The accuracy of classification is also analyzed based on different feature groups. The current feature groups, Texture, Color, Triangle, WDD, Nuclei, Light-Area, and Combined

Features, are tested in individual group and in combinations of groups for accuracy comparison.

For scoring the test image classifications, three different approaches are examined, Exact Classification, Off-By-One Classification, and Normal vs. CIN. The first approach is exact classification, meaning that if the class label automatically assigned to the test image (based on the algorithm above) is the same as the expert class label for the image, then the image is considered to be correctly labeled. Otherwise, the image is considered to be incorrectly labeled. The second scoring approach is an off by one scheme. If the predicted CIN grade level is only one value off as compared to the actual CIN grade, we considered it as correct prediction. For example, if CIN 1 was predicted as Normal or CIN 2, the result would be considered correct. If CIN 1 was predicted as CIN 3, the result would be considered incorrect. For the third approach, we considered the prediction incorrect when a normal stage was predicted as any CIN stages or vice versa.

7. EXPERIMENTATION RESULTS AND ANALYSIS

7.1. IMAGE-BASED CLASSIFICATION RESULT USING HORIZONTAL SEGMENTS

Using the horizontal segments for image-based classification, Table 7.1 and 7.2 below present classification results for the original 29 image data set. Table 7.1 shows exact class (actual image) classification results for SVM, LVQ, and Bayes classifiers for different feature combinations. Table 7.2 gives the off by one class label results for SVM, LVQ, and Bayes classifiers for different feature combinations. By evaluating the results from Table 7.1 and Table 7.2, between features groups, color features have a higher accuracy in general. Triangle feature produce the worst results. Any feature combination with triangle feature produce worse results compare to feature combination without triangle features. Between different classifiers, SVM classifier gives a higher accuracy over Bayes and LVQ classifiers in general. Due to this experimentation result, SVM classifier will be the only classifier to be experimented in vertical segmented classification section. For the exact classification, the best feature combination appears to be texture, color, triangle, and WDD, which are all the features developed at the time being. The accuracy for exactly classification is 65.52%. The off-by-one scoring scheme gives the best classification result of 96.55% for texture and color feature group combination.

7.2. IMAGE-BASED CLASSIFICATION RESULT USING VERTICAL SEGMENTS

Using the vertical segments for image-based classification, Table 7.3 below gives the classification results for the 62 image data set for the SVM classifiers for different feature combinations based on exact class labeling. Inspecting the results from Table 7.3,

there are several observations. First, the Off-By-One classification rates are as high as 100% for Nuclei, Light-Area, Combined, Texture feature group. Second, by adding the newly developed features, Nuclei, Light-Area, and Combined features, the classification accuracy improves significantly compare to feature combinations without them. Third, the normal vs. CIN scoring scheme gives a highest accuracy of 93.55% for a combination of color, nuclei, light-area, and combined features.

Table 7.1. Horizontal segment analysis for image-based classification using the original 29 image set for exact class label image-based classification results.

Exact Classification	Different Classifiers (%)		
	SVM	LVQ	Bayes
Texture, Color, Triangle, WDD	65.52	44.83	58.62
Texture	41.38	44.83	41.37
Color	51.72	44.83	58.62
Triangle	48.28	34.48	51.72
WDD	37.93	27.58	41.38
Texture, Color	51.72	31.03	55.17
Texture, Triangle	62.07	41.38	51.72
Texture, WDD	37.93	37.93	37.93
Color, Triangle	62.07	37.93	62.07
Color, WDD	58.62	41.38	58.62
Triangle, WDD	41.38	37.93	41.38
Texture, Color, Triangle	58.62	31.03	62.07
Texture, Color, WDD	62.07	41.38	55.17
Texture, Triangle, WDD	44.82	34.48	37.93
Color, Triangle, WDD	55.17	41.38	55.17

Table 7.2. Horizontal segment analysis for image-based classification using the original 29 image set for off-by-one image-based classification results.

Off-By-One Classification	Different Classifiers (%)		
	SVM	LVQ	Bayes
Texture, Color, Triangle, WDD	86.21	79.31	79.31
Texture	79.31	68.97	62.07
Color	93.10	82.76	93.10
Triangle	68.97	72.41	68.97
WDD	48.28	62.07	58.62
Texture, Color	96.55	62.07	82.76
Texture, Triangle	86.21	72.41	68.97
Texture, WDD	55.17	58.62	51.72
Color, Triangle	93.10	75.86	86.21
Color, WDD	79.31	79.31	79.31
Triangle, WDD	51.72	55.17	55.17
Texture, Color, Triangle	93.10	75.86	86.21
Texture, Color, WDD	82.76	75.86	75.86
Texture, Triangle, WDD	65.52	62.07	55.17
Color, Triangle, WDD	75.86	68.97	75.86

Table 7.3. Vertical segment analysis for image-based classification result using SVM classifier for 62 image data set.

Combined images (62)	Exact Classification (%)	Off-By-One Classification (%)	Normal vs. CIN Classification (%)
Texture, Color, Triangle, WDD	56.45	98.39	88.71
Texture	27.42	83.87	74.19
Color	46.77	95.16	87.10
Triangle	20.97	74.19	72.58
WDD	38.71	87.10	82.26
Texture, Color	51.61	95.16	85.48
Texture, Triangle	27.42	82.26	74.19
Texture, WDD	41.94	91.94	87.10

Table 7.3. Vertical segment analysis for image-based classification result using SVM classifier for 62 image data set (cont.).

Color, Triangle	35.48	93.55	85.48
Color, WDD	59.68	93.55	91.94
Triangle, WDD	41.94	80.65	82.26
Texture, Color, Triangle	51.61	95.16	88.71
Texture, Color, WDD	54.84	98.39	91.94
Texture, Triangle, WDD	50.00	93.55	91.94
Color, Triangle, WDD	53.23	93.55	88.71
Nuclei, Light-Area, Combined, Texture	66.13	100.00	90.32
Nuclei, Light-Area, Combined, Color	58.06	98.39	93.55
Nuclei, Light-Area, Combined, Triangle	56.45	95.16	90.32
Nuclei, Light-Area, Combined, WDD	58.06	96.77	88.71
Nuclei, Light-Area, Combined, Texture, Color, Triangle, WDD	72.58	98.39	93.55

8. CONCLUSION AND FUTURE SCOPE

In this thesis, the author has involved in most aspects of this research. The goal for this research is and will always be improvement of current algorithm and classification accuracy. The current algorithm gives the exact classification accuracy as high as 72.58% for a four-class classification problem, and off-by-one classification accuracy as high as 100% for a four-class classification problem. Ideally, NLM would like classification accuracy as high as possible. The current algorithm has been modified many times, by removing old unsatisfied feature groups, adding new feature groups, selecting the best suitable classifier, and examining each classification result in depth. There are many aspects that can be improved in the future, aspects such as, using different adaptive critic design as classifiers, using fuzzy logic algorithm for feature extraction and classification stages, performing error analysis for classification algorithm.

BIBLIOGRAPHY

[1] V. Kumar, A. Abbas, N. Fausto, and J. Aster, "Chapter 22 The Female Genital Tract, Figure 22-17 Spectrum of cervical intraepithelial neoplasia," *Robbins & Cotran Pathologic Basis of Disease,* p. 1020, 2009.

[2] R.J. Stanley, S. De, C. Lu, B. Cheng, "NLM Project Report: September 2012," technical report, National Library of Medicine, 2012.

[3] B. Cheng, S. Antani, R.J. Stanley, G.R. Thoma, "Graphical Image Classification Combining an Evolutionary Algorithm and Binary Particle Swarm Optimization", Proceedings of SPIE Electronic Imaging, San Francisco, California, Jan 2012, vol. 8297, pp.1-8.

[4] R.M.Haralick, K. Shanmugan, I. Dinstein, Textural Features for Image Classification, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 610-621, 1973.

[5] S. J. Keenan, J. Diamond, W. Glenn McCluggage, H. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton, "An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)," The Journal of Pathology, vol. 192, pp. 351-362, 2000.

[6] R.J. Stanley, W.V. Stoecker, & R.H. Moss, "A basis function feature-based approach for skin lesion discrimination in dermatology dermoscopy images," *Skin Research and Technology*, 14(4), 425-435, 2008

[7] J. Piper, E. Granum. On fully automatic feature measurement for banded chromosome classification. Cytometry 1989:10:242-255.

[8] R.J. Stanley, C. Lu, K. Banerjee, X. Pan, & P. Guo, "NLM Project Report: February 2013," technical report, National Library of Medicine, 2013.

[9] R.C. Gonzalez, and R. E. Woods. "Image Enhancement in the Frequency." *Digital Image Processing*. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2008. 147-215.

VITA

Cheng Lu was born in Shenyang in the province of Liaoning, China in 1987. He did his schooling at the QiaoLiang Elementary School (1994-2000), 95th Middle School (2001-2002), Rolla Junior High School (2003-2004) in Rolla, Missouri, and Rolla High School (2005-2007) before going to Missouri University of Science and Technology in Rolla, MO for his Bachelor of Science degree in Electrical Engineering from the Department of Electrical and Computer Engineering (2011). He stayed at Missouri University of Science and Technology, where he is expected to receive his Master of Science degree in Computer Engineering from the Department of Electrical and Computer Engineering in 2013.