University of Wollongong

Research Online

Faculty of Engineering and Information
Sciences - Papers: Part A

Faculty of Engineering and Information
Sciences

1-1-2012

Computer aided decision support system for cervical cancer classification

- Rahmadwati

University of Wollongong, rr516@uowmail.edu.au

Golshah Naghdy

University of Wollongong, golshah@uow.edu.au

Montserrat B. Ros

University of Wollongong, montse@uow.edu.au

Catherine Todd

University of Wollongong in Dubai, cath@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/eispapers>

Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Computer aided decision support system for cervical cancer classification

Abstract

Conventional analysis of a cervical histology image, such a pap smear or a biopsy sample, is performed by an expert pathologist manually. This involves inspecting the sample for cellular level abnormalities and determining the spread of the abnormalities. Cancer is graded based on the spread of the abnormal cells. This is a tedious, subjective and timeconsuming process with considerable variations in diagnosis between the experts. This paper presents a computer aided decision support system (CADSS) tool to help the pathologists in their examination of the cervical cancer biopsies. The main aim of the proposed CADSS system is to identify abnormalities and quantify cancer grading in a systematic and repeatable manner. The paper proposes three different methods which presents and compares the results using 475 images of cervical biopsies which include normal, three stages of pre cancer, and malignant cases. This paper will explore various components of an effective CADSS; image acquisition, pre-processing, segmentation, feature extraction, classification, grading and disease identification. Cervical histological images are captured using a digital microscope. The images are captured in sufficient resolution to retain enough information for effective classification. Histology images of cervical biopsies consist of three major sections; background, stroma and squamous epithelium. Most diagnostic information are contained within the epithelium region. This paper will present two levels of segmentations; global (macro) and local (micro). At the global level the squamous epithelium is separated from the background and stroma. At the local or cellular level, the nuclei and cytoplasm are segmented for further analysis. Image features that influence the pathologists' decision during the analysis and classification of a cervical biopsy are the nuclei's shape and spread; the ratio of the areas of nuclei and cytoplasm as well as the texture and spread of the abnormalities. Similar features are extracted towards the automated classification process. This paper will present various feature extraction methods including colour, shape and texture using Gabor wavelet as well as various quantative metrics. Generated features are used to classify cells or regions into normal and abnormal categories. Following the classification process, the cancer is graded based on the spread of the abnormal cells. This paper will present the results of the grading process with five stages of the cancer spectrum.

Keywords

computer, cancer, classification, support, decision, system, cervical, aided

Disciplines

Engineering | Science and Technology Studies

Publication Details

Rahmadwati, G. Naghdy, M. B. Ros & C. Todd, "Computer aided decision support system for cervical cancer classification," The Proceedings of SPIE: Applications of Digital Image Processing XXXV, vol. 8499, (2012) pp. 1-13, 2012.

COMPUTER AIDED DECISION SUPPORT SYSTEM FOR CERVICAL CANCER CLASSIFICATION

Rahmadwati, Golshah Naghdy, Montserrat Ros, Catherine Todd*

School of Electrical, Computer and Telecommunication Engineering University of Wollongong
*Faculty of Computer Science and Engineering, University of Wollongong in Dubai, Dubai, United Arab Emirates,

rr516@uowmail.edu.au, golshah@uow.edu.au, montse@uow.edu.au, CatherineTodd@uowdubai.ac.ae

ABSTRACT

Conventional analysis of a cervical histology image, such a pap smear or a biopsy sample, is performed by an expert pathologist manually. This involves inspecting the sample for cellular level abnormalities and determining the spread of the abnormalities. Cancer is graded based on the spread of the abnormal cells. This is a tedious, subjective and time-consuming process with considerable variations in diagnosis between the experts. This paper presents a computer aided decision support system (CADSS) tool to help the pathologists in their examination of the cervical cancer biopsies. The main aim of the proposed CADSS system is to identify abnormalities and quantify cancer grading in a systematic and repeatable manner. The paper proposes three different methods which presents and compares the results using 475 images of cervical biopsies which include normal, three stages of pre cancer, and malignant cases.

This paper will explore various components of an effective CADSS; image acquisition, pre-processing, segmentation, feature extraction, classification, grading and disease identification. Cervical histological images are captured using a digital microscope. The images are captured in sufficient resolution to retain enough information for effective classification. Histology images of cervical biopsies consist of three major sections; background, stroma and squamous epithelium. Most diagnostic information are contained within the epithelium region. This paper will present two levels of segmentations; global (macro) and local (micro). At the global level the squamous epithelium is separated from the background and stroma. At the local or cellular level, the nuclei and cytoplasm are segmented for further analysis. Image features that influence the pathologists' decision during the analysis and classification of a cervical biopsy are the nuclei's shape and spread; the ratio of the areas of nuclei and cytoplasm as well as the texture and spread of the abnormalities. Similar features are extracted towards the automated classification process. This paper will present various feature extraction methods including colour, shape and texture using Gabor wavelet as well as various quantitative metrics. Generated features are used to classify cells or regions into normal and abnormal categories. Following the classification process, the cancer is graded based on the spread of the abnormal cells. This paper will present the results of the grading process with five stages of the cancer spectrum.

Keyword: computer aided decision support system, cervical cancer, histology image, K-means clustering, Gabor wavelet, graph cut segmentation

1. INTRODUCTION

Currently, slide examination of cervical histology images is performed manually by the pathologist. An expert pathologist views a glass slide under the microscope and analyses the whole image in order to identify the region of interest (ROI). Pathologists examine the slide based on their experiences, and individual pathologists can interpret the slide differently [1, 2]. The examination of the slide is therefore influenced by subjectivity factors [2]. Furthermore, a problem arises for laboratory technicians when the experts are not available to examine the slides [2].

Cervical cancer is the second highest cause of cancer death among women in the world [3]. The introduction of the vaccination in 2006 has helped in the prevention of infection against human papillomavirus (HPV) type 16 and 18, which are linked to 70% of cervical cancer cases [4]. While the efficacy of HPV vaccine is promising, close to 95% [4-5], researchers cannot predict the time period that the vaccine remains effective against the HPV virus or know how to help the patients who already have HPV [6]. There is no guarantee that HPV vaccine will protect the recipient indefinitely and experts recommend regular screening even for those who have been vaccinated [7]. There are still considerably high numbers of cases of cervical cancer resulting in death. Cervical cancer incidences and fatalities are particularly common in developing countries due to lack of adequate testing, laboratory equipment, technology and technician skills [8]. It is therefore of benefit to develop a system which assists pathologists in the diagnostic process and reduces the subjectivity associated with the diagnosis. Therefore, a computer aided decision support system (CADSS) for cervical cancer classification is proposed. The development of the proposed CADSS requires the integration of the research areas including mathematics, digital image processing, physics, computer vision and statistics.

Studies have documented computer aided diagnosis (CAD) to detect and classify abnormalities of cells within histology images [2, 9-14]. One of the processes in CADSS is image segmentation. The purpose of segmentation is to partition the ROI. The segmentation methods applied in the research for ROI identification within the histopathology image are edge detection [15-16], active contours [17-19], threshold-based approaches [11,20], K-Means clustering [1,13], Local Gaussian (LG) and Gaussian Mixture Model (GMM) [21], and application of the co-occurrence matrix [2].

This paper introduces a systematic, quantitative, accurate, repeatable and objective CADSS to help practitioners in increasing the sensitivity and specificity of pathology examinations for cervical cancer. This study uses cervical histology images as input to the CADSS system. A typical CADSS consists of sequences of image processing and analysis tools including image acquisition, pre-processing, segmentation, feature extraction, classification, gradation and disease identification [22-24].

2. IMAGE ACQUISITION

The histological data that are used in this study are taken from cervical biopsy samples. A cervical biopsy is recommended when abnormal tissues or suspicious lesions are discovered during colposcopy.

In collaboration with the staff in the pathology anatomy laboratory, Saiful Anwar hospital in Indonesia, the records of biopsy data from the patients with suspected cervical cancer were searched for suitable representative data from a range of cancer grades. The data consists of normal, Cervical Intraepithelial Neoplasia (CIN) 1, CIN2, CIN3 and malignant as diagnosed by the pathologists.

The slides were captured by a digital microscope provided by Brawijaya University, Indonesia. The images were captured using a charge-coupled device (CCD) camera integrated with the microscope. The cervical histology slides were taken with a digital microscope at 400 level magnifications and the spatial resolution of 4080x3072 pixels. The hospital's pathologist classified 475 histological images consisting of 60 normal, 70 CIN1, 50 CIN2, 50 CIN3 of pre cancer, and 245 malignant cases. Examples of the cervical histology images are shown in Figure 1.

Figure 1 Examples of the cervical histology images

3. METHODS

This paper presents the results of three different methods for the diagnostic analysis of the cervical histology images. In this section some of the algorithms used in these methods are discussed.

3.1 Image Pre-processing

Generally, noise in histology images can be attributed to debris of nuclei, blood cells and artefact stains in the background [25]. Pre-processing is required to reduce the noise and improve the quality of the image in order to determine the ROI. Pre-processing includes filtering to reduce noise and to enhance the image so that the result is an improvement on the original [26]. Many methods, such as thresholding and adaptive filtering, have been proposed to eliminate and filter noise [22]. The filtering method that is used in this study is the median filter. A number of studies apply the median filter method in the pre-processing process to medical images [13, 27-28]. The median filter sorts the values (according to the brightness or intensity) of each neighbouring pixel in ascending order [1, 13]. Subsequently, the median value of this ordered sequence is selected. Median filter is suitable for removing random and periodic noise.

3.2 Image Segmentation

Segmentation refers to the separation of a histology image into distinct regions. The most obvious components of a cervical histology image are stroma, squamous epithelium and the background, as shown in Figure 2. The important diagnostic information such as sizes, shapes, textures and densities of the nuclei are contained within the squamous epithelium region. At the cellular level, the nuclei inside the epithelium region should be segmented and analysed for diagnostic processes. In this study, the segmentation of the cervical histology images is a scalable process which is capable of segmentation at coarse and fine scales (global and local). In the following sub-section the segmentation methods that are used in this study are introduced.

Figure 2. Cervical histology image with its components

3.2.1 K-means clustering for local segmentation

K-means clustering is one of the simplest methods for segmentation. The algorithm of K-means clustering is used to segment an image based on its features such as colour and texture [29]. The process of clustering is achieved by measuring the sum of distance between the data and the cluster centroid [29].

The general algorithm of K-means clustering is as follows [29]:

- The number of clusters are predefined and the cluster centers are selected.
- The distances between the sample and the cluster centroids are calculated to determine each sample's closest cluster and to generate new partition.
- Based on the distance (commonly Euclidian distance) of the pixel similarity, the pixel is clustered by application of Equation 1.

$$c^{(i)} := \arg \min_j \|x^{(i)} - \mu_j\|^2 \quad (1)$$

Where $c^{(i)}$ is a pixel cluster, data point $x^{(i)}$ and cluster center μ_j .

- The processes are repeated until a distance convergence criterion is met.
- The calculation of each new formed cluster is achieved by application of Equation 2

$$\mu_i := \frac{\sum_{i=1}^m 1\{c(i)=j\} x^{(i)}}{\sum_{i=1}^m 1\{c(i)=j\}} \quad (2)$$

Where μ_i is new cluster

The principle of K-means is that the clustering is defined by the region's mean intensity value. The goal of this method is to find the partition of the n data points into the k cluster by minimising the sum of squared distance between the data and the centroid that is presented in Equation 3 [29].

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (3)$$

Where S_i = cluster, $i=1, 2, 3, \dots, k$ and k is the number of cluster, i = iteration for all the pixels images, j = iteration to the all centroids and μ_i = the new cluster, x_j = data point.

3.2.2 Graph cut segmentation for global segmentation

The graph cut is a process that involves the partitioning of a directed or undirected graph into disjoint sets [30]. Each pixel contained within the image is considered a vertex or node of the graph itself [31]. The similarity between two pixels represents the weight of the edge between the two corresponding vertices. In graph theory, the nodes of the graph represent points in feature space.

A weighted undirected graph is represented by the following Equation

$$G=(V,E) \quad (4)$$

Where G is a weighted undirected graph that represents the set points in an arbitrary feature space; V is a set of graph nodes (vertices); and E is a similarity between two pixels represent the weight of edge between the two corresponding vertices [32].

The first step to apply in the graph cut segmentation process is to translate the image into graphical representation. Within this step, each pixel is presented as a node and the connecting line between the nodes is called an edge.

A graph can be divided into two disjoint sets A and B . The level of dissimilarity between A and B can be calculated by a cut in Equation 5.

$$\text{cut}(A,B) = \sum_{i \in A, j \in B} w(i,j) \quad (5)$$

Where $w_{i,j}$ denote the weight on each edge that refers to the similarity between the nodes i and j .

By minimising the cut value, the optimal graph partition will be obtained [33-34]. The problem is finding the minimum cut [32]. Shi and Malik proposed a new method for measuring the disassociation called the normalised cut that is defined

as a fraction of the total edge connections to all the nodes in the graph [32]. Equation 6 represents the normalised cut (N_cut) and Equation 7 expresses the total of connections in the graph [32].

$$N_cut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \quad (6)$$

$$assoc(A,V) = \sum_{i \in A, j \in V} w(i,j) \quad (7)$$

Where, $assoc(A, V)$ is defined as the total connections from nodes A to all nodes in the graph and $assoc(B, V)$ is similarly defined. An important part of the graph cut segmentation is the determination of the minimum cut parameter to obtain the optimal partition graph [32].

The total normalised association within groups for given partition is expressed in Equation 8 which reflects how closely on average nodes within the cluster are related to each other [32].

$$Nassoc(A,B) = \frac{assoc(A,A)}{assoc(A,V)} + \frac{assoc(B,B)}{assoc(B,V)} \quad (8)$$

Where $assoc(A, A)$ and $assoc(B, B)$ are the total weights of edges connecting nodes within A and B, respectively [32].

Assume D be the total connection from the node i to all other nodes that is an $n \times n$ diagonal matrix with entries

$$D(i,i) = \sum_j w(i,j) \quad (9)$$

And W is an $n \times n$ symmetrical matrix

$$W(i,j) = w(i,j) \quad (10)$$

The minimisation cut is expressed in Equation 11 [31].

$$\min_{A,B} N_cut(A,B) = \min_y \frac{y^T(D-W)y}{y^T D y} \quad (11)$$

$$\text{With the condition } y^T D y = 1 \quad (12)$$

Equation 11 is also called the Rayleigh quotient [32]. If y is relaxed to take on real values, the Equation 11 can be minimised by solving the generalized Eigen value system as Equation 13 [32].

$$(D-W)y = \lambda D y \quad (13)$$

It is important to note that if

$$(D-W)1 = 0 \quad (14)$$

Then the first eigenvector is $y=1$ with Eigen value of zero [32]. Thus, the second smallest eigenvector y provides solution in the normalised cut problem.

There are several steps in order to segment an image based on the grouping algorithm in normalised cut method:

1. Given an image, convert into weighted graph as Equation 4. Subsequently, set the weight on the edge connecting two nodes.
2. Solve Equation 13 for Eigen vectors with the smallest Eigen value.
3. In order to bipartition the graph, the Eigen vector is used with the second smallest Eigen value.
4. The partition is performed iteratively (repeated) until all parts are segmented.

3.2.3 Morphology operation

The morphological operation aims to improve the quality of an image after the segmentation process [33]. The principle of morphological operation is based on the shape and size of an object named structural element. The structural element is modified by using a mathematical morphology operation. The shape and size of the structural elements are adjusted according to the segmentation or filtering task [33]. The size of the structural element is smaller than the input image; typically the size of the structure is 3×3 , 4×4 or 5×5 pixels [34]. The shape of structural elements varies and includes shapes such as rectangles, disks, diamonds, balls, lines, squares and octagons [22]. The basic concepts of the morphology operation are dilation and erosion that is explained below:

3.2.3.1 Dilation

Dilation is defined as an operation to increase the size of an object [22]. Generally speaking, dilation is used for filling the gaps or holes in the image and for connecting the separated objects [22].

3.2.3.2 Erosion

Erosion is a method of morphology image processing that erodes an object in an image [22]. The object size is decreased and the small peninsula in the object is removed by subtracting the objects with a distance less than the structural element [22].

3.3 Feature Extraction

The partitioning and characterisation of cell images in the cytology and histopathology of tissue cancer are the main focal points of the CADSS system. It is expected that the feature extraction process will extract appropriate characteristics of the input data with the purpose of reducing the complexity of the representation. A CADSS for grading diseases based on histology images depends upon capturing the variation of cell structure (at the cellular level), and

changes in the cell-spread at the tissue level [7]. This study explores various methods for feature extraction such as texture features and morphological features.

3.3.1 Gabor Wavelet

Gabor Wavelet is used in multi channel filtering to localise an image into spatial and frequency domains in order to analyse the localised texture characteristics of the image. Gabor wavelet is designed through a combination of typical frequencies and orientations. Gabor wavelet filter banks are designed through certain procedures to calculate the filter parameters. Each filter produces one element of a feature vector in high dimensionality [35]. It is important to select the appropriate filter parameters in order to characterise textures embedded in the image for the segmentation process [14]. The finer detection of textures is obtained by applying the filters with higher frequencies [35]. In essence, a Gabor function is a complex sinusoid modulated by a Gaussian function. The Gabor wavelet bank is a multi-channel filter using specific spatial frequencies and orientations of a sinusoid and the standard deviation of Gaussian along x and y directions (σ_x and σ_y). The Gabor elementary function is represented by Equation 15.

$$h(x,y)=g(x',y')\exp[j2\pi(U_x+V_y)] \quad (15)$$

A two -dimensional Gaussian function is defined as Equation 16

$$g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y} \exp\left\{\frac{1}{2}\left[\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right]\right\} \quad (16)$$

where $x'=x \cos \theta + y \sin \theta$ and $y'=-x \sin \theta + y \cos \theta$.

From Equation 20 and 21, the Gabor elementary function can be rewritten as Equation 17.

$$g(x,y)=\frac{1}{2\pi} \exp\left[\frac{x^2+y^2}{\sigma_x^2+\sigma_y^2}\right] \exp[j2\pi u_0(x \cos \theta + y \sin \theta)] \quad (17)$$

u_0 is the frequency plane wave (centre frequency of the sinusoid), where σ_x and σ_y are the spread of the Gaussian (standard deviation of Gaussian envelope) in x and y directions respectively. Assuming σ_x and σ_y are the same, and θ is the orientation of the sinusoid. The output of Gabor wavelet is transformed using a non linear function. Subsequently, the local energy is computed which results in a feature vector at each pixel location.

3.3.2 Morphological Calculation

In this research, the morphological calculation is used to extract the features of nuclei and cytoplasm. The morphological calculations include nuclei area, nucleus to cytoplasm ratio (N/C ratio), shape factor, compactness and triangle area of Delaunay Triangulation (DT). The information concerning size and shape of the nuclei and cytoplasm are described by the morphological features calculated as follows:

3.3.2.1 Nuclei area

The measurement of nuclei area is performed by counting the number of pixels in a nucleus or calculating the diameter of a nucleus. This feature is designed to analyse the characteristics of individual nuclei in the image block.

3.3.2.2 Nucleus to Cytoplasm ratio (N/C ratio)

One of the criteria to determine the abnormality in cervical histology is the ratio of nucleus area to cytoplasm area [36]. Generally, abnormal cells have higher nuclei to cytoplasm (N/C) ratios, which may approach 1:1, while normal cells have much lower ratios; between 1:4 and 1:6 [37]. The N/C ratio is measured using Equation 18.

$$N/C \text{ ratio} = \frac{\text{nuclei area}}{\text{cytoplasm area}} \quad (18)$$

3.3.2.3 Shape factor

One of the features used to identify abnormal nuclei is shape. The nuclei may be round, tadpole, bizarre, oval or caudate [39]. The abnormal nuclei shape is irregular, loss of roundness, indentation and lobes [36]. In comparison, normal nuclei have a smooth nuclear envelope, circular shape (round nucleus), small chromocenter and regular [36]. The shape factor is used to measure or assess the shape of the nuclei. The shape factor is measured by roundness factor. Equation 19 represents the roundness factor of the nuclei [38].

$$\text{Roundness} = \frac{4\pi \text{Area}}{\text{perimeter}^2} \quad (19)$$

3.3.2.4 Compactness

Compactness is the ratio of the square of the perimeter to the area of the nuclei [33]. The compactness is used measure the contour complexity versus closed area [42]. In this work, the compactness is calculated using Equation 20 [33, 42].

$$\text{Compactness} = \frac{\text{perimeter}^2}{\text{Area}} \quad \text{where the perimeter} = 2\pi \quad (20)$$

3.3.2.5 Area of Delaunay Triangulation

A triangular area in DT is used in this research. This feature measures the nuclei density of a region which is an indication of abnormalities. The centre of each nucleus is assigned to the vertex point. DT is established by connecting at least three nuclei so that none of the three points are inside the circum circle of any other triangles. The area of the

triangle measures the distance between nuclei in the image block. The mean area of triangles in DT is calculated for the extraction of the feature. The area of triangle in DT is shown in Figure 3 which marked in blue line.

Figure 3 The area triangle of DT

3.4 Classification Tree

Classification tree is one of the categorising methods using the concept of a tree structure to classify the features derived from the image. In the classification tree the features from data samples will be processed by the logical IF-THEN condition. This will split the feature variables into non-overlapping regions, in which the value of the dependent variable will be predicted. The classification tree is processed using MATLAB software.

3.5 Grading and Disease Identification

The next step is the grading process that defines the stage of the disease based on the standardised nomenclature of cervical cancer classification by Richart [43]. The numbers of normal and abnormal cells are quantified to determine their class. Classification is based on the following algorithm where the pathologists' view and diagnostic clues in linguistic terms for grading the disease, is translated into a quantitative measurement. The interpretation of diagnostic measurements is designed to correspond as closely as possible to the interpretation of expert pathologists.

The classification is defined according to the following rules:

1. Normal: the ratio between the number abnormal cells to the total number of cells is less than 1/10.
2. CIN 1: the ratio between the number abnormal cells to the total number of the cells is not more than 1/3.
3. CIN 2: the ratio between the number abnormal cells to the total number of the cells is more than 1/3 but not more than 2/3.
4. CIN 3: the ratio between the number abnormal cells to the total number of the cells is more than 2/3 but not more than 4/5.
5. Malignant: the ratio between the number abnormal cells to the total number of the cells is between 4/5 and close to 1.

4. EXPERIMENTATION

In this section the experimental processes for the three proposed methods are described.

4.1. Method 1; K-means clustering based on the colour

The processes of CADSS for cervical cancer using K-means clustering based on the colour are shown in Figure 4. In the first stage, the cervical histology images are pre-processed using Median filter. The images are then segments into their constituent elements using color based K-means clustering method. There are four dominant colours: white for the background, dark purple for the nuclei, dark pink for the cytoplasm, light pink for stroma [1, 13, 7]. The colours are the outcome of the Hematoxylin and Eosin (H&E) staining in the cervical histology slide for ease of analysis. According to the dominant colours of cervical histology images, K-means clustering is applied for the segmentation of the histology image into their regions.

Figure 4. K-means clustering based on the colour

Prior to the segmentation of the cervical histology image using K-means clustering, the numbers of clusters are defined, there are four clusters. K-means clustering utilises the Euclidian distance metric of colour to classify the regions. The K-means clustering process used in the present research is outlined in the following steps.

1. The filtered image is converted to the International Commission on Illumination(CIE) (L^*a^*b) colour space. The segmentation method used in this research is based on the colours converted from the red, green, blue (RGB) format to the CIE L^*a^*b format. The 'L' component describes the brightness of the colour [22]. Component 'a' represents the colour placed along the red and green axis and 'b' indicates the colour placed along the blue and yellow axis [22]. The reason for converting the RGB colour to the CIE is reduced dimensionality as in CIE format we only need to consider a^* and b^* for color analysis.
2. In this process, colour segmentation of cervical histology regions is based on the Euclidian distance in which the distance is measured for each defined colour. The locations of the initial cluster centroid are initialised and the colour distances are quantified by the distance metric. Pixels with the same distance are clustered in the same class.
3. Each region is obtained by segmenting the image into four colours which later is displayed in greyscale form. The result of applying K-means clustering to a cervical histology image is shown in Figure 5.

Figure 5 (a) Original Image (b) Result of K-means clustering based on the colour

The black colour refers to the nuclei, the white colour presents the cytoplasm, the dark grey refers to the background and the light grey shows the stroma.

In stage three, the nuclei and the cytoplasm are separated. The morphological operation aims to improve the quality of an image after the segmentation process. After the nuclei and cytoplasm images are segmented and displayed in different images, the nuclei and cytoplasm images are processed using the matching method; that is aimed at finding the correlated location of the nuclei and cytoplasm. Subsequently, the features of nuclei and cytoplasm are extracted based on the morphological features calculation including area, perimeter, diameter, shape factor and compactness. In the training process, 75 samples of normal and 75 samples of abnormal nuclei are taken in order to identify their characteristics. The characteristics of normal and abnormal nuclei are obtained in order to find the threshold values for each feature to be used in the testing process. The properties of normal and abnormal cells are obtained based on the classifier shown on the Table 1.

Table 1 The features of nuclei

Type	N/C	Shape factor	Compactness	Diameter (pixel)
Normal	0.0265± 0.006	0.891	12.30	13.26±0.0045
Abnormal	0.2246±0.002	0.735	13.56	18.49±0.00593

According to Table 1, the abnormal cells have the greatest values in the N/C ratio, compactness and diameter whilst the normal cells have the greatest shape factor value. In the normal case, the cytoplasm is apparently very large whereas the nuclei are classified as small. The system can distinguish between the normal and abnormal cells based on the features above.

4.2. Method 2; Gabor wavelet for texture classification

Figure 6 The block diagram of Gabor wavelet in CADSS

Figure 6 presents the stages in CADSS using Gabor wavelet. First, images are converted into greyscale with a spatial resolution of 4080 x 3072 pixels, with 8-bit pixel depth. The images are further processed using Gabor wavelet banks with twenty four filters; six frequencies and four orientations.

The steps in the application of the Gabor wavelet to the images are further described below.

1. Apply the Gabor wavelet bank with six frequencies and four orientations for each pixel. The smallest frequency is $32\sqrt{2}$ and the largest is $1024\sqrt{2}$. The orientations of the Gabor wavelet used in this research are 0, 45, 90 and 135 degrees.
2. Compare the features (the magnitude of the Gabor wavelet bank) of each pixel with the templates obtained during the training process
3. Tag each pixel with one of the five categories according to the nearest neighbour process.
4. K-means clustering based on the colour is applied to partitioning the image that resulted from Gabor wavelet into their component based on the colour.
5. Calculate the ratio of abnormal cells to total number of cells.
6. Classify the image into normal, malignant or pre cancer in accordance to how a pathologist would examine the slide.

According to the nearest neighbourhood assessment, each pixel is tagged (visualisation) with a different colour, using the following colour coding: blue is background, yellow is basal cell, white is stroma, green as normal cell and red is an abnormal cell. A result of the Gabor wavelet feature extraction is presented in Figure 7.

Figure 7 (a) Original Image (b) Output image of Gabor wavelet

Figure 7 (b) shows an output image that has been extracted from a Gabor wavelet. K-means clustering is used to clustering the output of Gabor wavelet into separated image. The following section describes the K-mean clustering result. An image revealing normal and abnormal nuclei is shown in Figure 8.

Figure 8. The result of K-means clustering based on the colour
(a) background, (b) basal and stroma, (c) abnormal cell and
(d) normal cell

At the end of the previous stage, K means clustering segments the output image based on the colour. As a result, the normal and abnormal nuclei are separated. The system recognises the normal and abnormal nuclei, and counts the number of those nuclei. In accordance with how a pathologist examines the slide, the image is classified into normal, pre cancer (CIN1, CIN 2, CIN3) and malignant.

4.3. Method 3: Hybrid graph cut and colour segmentation

Figure 9 Block diagram of hybrid graph cut and colour segmentation in CADSS

Figure 9 presents the hybrid graph cut and colour segmentation. There are two segmentation approaches in this method: global and local. In the global approach graph cut segmentation is used to separate the squamous epithelium from

background and stroma. In order to obtain the optimal partition the normalised cut is applied that measures the total dissimilarity between different regions and the total similarities between regions [32]. In the local approach the colour segmentation method is used to extract the nuclei. First, the squamous epithelium is divided into blocks of 500 x 500 pixels. The size of the block is chosen to optimise the two processes; to ensure that the area of image block is large enough to include several nuclei and is small enough for a fine classification. Each block is segmented using the colour segmentation method. The purpose of this process is to analyse the cellular level in each block based on the morphology characteristics of the nuclei, then classify it as a normal or abnormal region. The nuclei area and other features are extracted by DT [11]. The DT is built by connecting every three cells into one triangle. The mean area of nuclei and the mean area of triangles in DT are used for extracting the morphology characteristics of nuclei. The input to the classification training process is the features of normal and abnormal regions. In the training process, 100 samples of normal and abnormal regions are taken. There are differences in appearance between the normal and the abnormal regions in the squamous epithelium layer. The density of nuclei in the abnormal region tends to be higher than the normal region. Samples are taken from the normal and abnormal regions in order to obtain the threshold value of normal and abnormal region quantitatively based on a pathologist's analysis. The samples of normal and abnormal region are shown in Figure 10.

Figure 10. The samples of image block (a) abnormal region
(b) normal region

The classification tree will predict the class to which block belongs. The result of a classification tree is depicted in Figure 11. It shows the mean values of the nuclei area and the triangle of DT which distinguish the normal and abnormal blocks. This classification tree is computed using MATLAB software.

Figure 11. The classification tree

The testing process is applied to the epithelium layer by sliding a window size of 500 x 500 pixels through the layer and analysing the contents of each window in order to classify it as normal or abnormal. Whilst researchers have applied the sliding block for feature extraction [42], this work is unique due to the use of this technique to analyse cervical cancer at the cellular level. The sliding block examines the nuclei at close proximity and determines the abnormal or normal segments. The testing phase aims to analyse the nuclei and their spread in detail. The sliding block moves from the left top corner toward to the right in horizontal and vertical direction. The blocks move every ten pixels until all squamous epithelium regions have been analysed. The result of local approach is shown in Figure 12. Each block is labelled into normal and abnormal blocks, with green representing the normal block and red representing the abnormal block. Figure 12 presents the original image; the graph cut segmentation (global approach) and the result of local approach. The total numbers of normal and abnormal regions are calculated. The ratios between abnormal and the total numbers of regions are used to classify the cervical histology images into normal, CIN1, CIN2, CIN3 and malignant categories.

5. RESULTS AND DISCUSSION

The result of three methods: K-means clustering, Gabor wavelet, and hybrid graph cut and colour segmentation are described in this section. The performance of each method is presented in the following tables. The system classifies the images into normal, CIN 1, CIN 2, CIN 3 or malignant. All images were classified according to the algorithms mentioned in the grading and classification sections. The confusion matrix of K-means clustering method is shown in Table 2.

Figure 12 The result of hybrid global and local approaches

Table 2 The result of the K-means clustering method

Category	Normal	CIN 1	CIN 2	CIN 3	Malignant	Total
Normal	48	9	3			60
CIN 1		53	15	2		70
CIN 2		2	43	5		50
CIN 3		2	3	45		50
Malignant			8	18	219	245

The confusion matrix of using twenty four Gabor wavelet is shown in Table 3

Table 3 Confusion matrix using twenty four Gabor wavelet

Category	Normal	CIN 1	CIN 2	CIN 3	Malignant	Total
Normal	52	7	1			60
CIN 1		60	9	1		70
CIN 2		3	41	6		50
CIN3		1	6	42	1	50
Malignant			3	23	219	245

The confusion matrix using hybrid graph cut and colour segmentation is presented in Table 4.

Table 4 Confusion matrix of graph cut and colour segmentation

Category	Normal	CIN1	CIN2	CIN3	Malignant	Total
Normal	58	2				60
CIN1		62	4	2		70
CIN2	1	1	47	1		50
CIN3		2	1	40	7	50
Malignant				6	239	245

A comparison of the performance of the three systems, including sensitivity, specificity, false positives (FP) and false negatives (FN) is displayed in Table 5.

A total of 475 cervical histology images were processed for cervical cancer diagnosis. Three methods were applied to analyse the images: K-means clustering based on colour, Gabor wavelet, and a hybrid of graph cut and colour segmentation. Comparing the results of these three methods, the sensitivity of K-means clustering and Gabor wavelet were 100%, whereas the hybrid of graph cuts and colour segmentation was 99.75%. Using a hybrid of graph cuts and colour segmentation, one case of CIN 2 was wrongly diagnosed as a normal case.

Table 5. The comparison result of CAD using K-means, Gabor wavelet and Hybrid Graph cut and colour segmentation method.

	K-means (%)	Gabor wavelet (%)	Graph cut and Colour (%)
Specificity (TN)	80	87	97
Sensitivity(TP)	100	100	99.75
False Positive (FP)	20	13	3.3
False Negative (FN)	0	0	0.24

It can be seen that the FP rate for K-means clustering was 20%, Gabor wavelet 13% and the FP rate for the hybrid of graph cuts and colour segmentation was 3.3%. Therefore, the FP rate of K-means clustering is the highest, whereas the hybrid of the graph cut and colour segmentation has the highest rate of FN. However, it is important to note that in medical diagnosis, the consequences of FN are far more severe than FP. The sensitivity of the hybrid of graph cuts and the colour segmentation method is 0.25% lower than other methods. The fact that a hybrid of a graph cut and colour segmentation method accurately classified 97% of the normal cases means that it outperforms the other two methods in specificity. The hybrid of graph cuts and colour segmentation performs less effectively in sensitivity and FN. However, it outperforms the other two in terms of minimising FP and maximising specificity. In this case, the hybrid of graph cuts and a colour segmentation system more accurately identifies normal conditions than other methods.

6.CONCLUSION

From the analysis of the performance of the segmentation and classification processes, the Gabor wavelet, and the hybrid graph cut and colour segmentation perform better than K-means clustering. Both of those methods segment the cervical histology in fine detail and provide more accurate results. The selection of filters influences the capability of the Gabor wavelet. Furthermore, in this case, the fine textures that are more visible at higher frequencies carry most of the important diagnostic information. Therefore Gabor wavelet tuned to higher frequencies capture essential diagnostic features. On the other hand, in the hybrid graph cut and the colour segmentation, the analysis shows that the sliding blocks reveal finer image classification when moved in steps of ten pixels.

According to the classification and grading results, the hybrid graph cut and the colour segmentation performs better than the others. It can be seen that the hybrid graph cut and the colour segmentation method results in the highest sensitivity and specificity whereas its FP is the lowest. However, it has a small number of FN which influences the diagnostic results.

References :

- [1] Rahmadwati, G. Naghdy, M. Ros, and C. Todd, "Morphological Characteristics of Cervical Cells for Cervical Cancer Diagnosis," *Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science*, vol. 145, pp. 235-243, 2012.
- [2] Y. Wang, R. Turner, D. Crookes, J. Diamond, and P. Hamilton, "Investigation of Methodologies for the Segmentation of Squamous Epithelium from Cervical Histological Virtual Slides," *International Machine Vision and Image Processing Conference* pp. 83-90, 2007.
- [3] X. Castellsagué, S. de Sanjosé, T. Aguado, K.S. Louie, J. M. L. Bruni, M. Diaz, K. Irwin, M. Gacic, O. Beauvais, G. Albero, E. Ferrer, S. Byrne, and F. X. Bosch, "HPV and Cervical Cancer in the World.2007 Report," *Vaccine, Report WHO/ICO Information Centre of HPV and Cervical Cancer (HPV Information Centre)*, vol. 25, 2007.
- [4] C.Washam, "Targeting Teens and Adolescents for HPV Vaccine Could Draw Fire," *Journal of the National Cancer Institute*, vol. 97, 2005.
- [5] Paavonen J, Bosch FX, Naud P, Salmerón J, Wheeler CM, Chow SN, Apter DL, Kitchener HC, Castellsague X, de Carvalho NS, Skinner SR, Harper DM, Hedrick JA, Jaisamrarn U, Limson GA, Dionne M, Quint W, Spiessens B, Peeters P, Struyf F, Wieting SL, Lehtinen MO, Dubin G; HPV PATRICIA study group., "Efficacy of a prophylactic adjuvanted bivalent L1 virus-like-particle vaccine against infection with human papillomavirus types 16 and 18 in young women: an interim analysis of a phase III double-blind, randomised controlled trial.," *Lancet*. 2007 vol. 369(9580). pp.:2161-70, 2007 Jun 30;.
- [6] Henderson, M, "Vaccine Could Wipe out Deadly Cervical Cancer.," *Alliance for Microbicide Development Weekly News Digest* 6(4). 2005.
- [7] Y. Wang, "Computer Assisted Diagnosis of Cervical Intraepithelial Neoplasia(CIN) Using Histological Virtual Slide," in *School of Electronics, Electrical engineering and Computer Science*: Queen University Belfast, 2008.
- [8] World Health Organisation, "Cervical Cancer Screening in Developing Countries," *Report of a WHO consultation*, 2002.

- [9] S.Doyle, M.Hwang, K.Shah, A.Madabhushi, J.Tomasezweski, and M.Feldman, "Automated Grading of Prostate Cancer Using Architectural Images Features," *Proc. IEEE International Symposium on Biomedical Imaging*, 2007.
- [10] Doyle.S, M.Hwang, K.Shah, A.Madabhushi, J.Tomasezweski and M.Feldman, "Automated Grading of Breast Cancer Histopathology Using Spectral Clustering with Textural and Architectural Image Features," *Proc. IEEE International Symposium on Biomedical Imaging*, 2008.
- [11] S. J. Keenan, J. Diamond, Glenn McCluggage, W. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton, "An Automated Machine Vision system for Histological Grading of Cervical Intraepithelial Neoplasia(CIN)," *Journal of Pathology*, vol. 192, pp. 351-362, 2000.
- [12] M. Jondet, R. Agoli-Agbo, and L. Dehennin, "Automatic Measurement of Epithelium differentiation and Classification of Cervical Intraepithelial Neoplasia by Computerized Image Analysis," *Diagnostic Pathology*, 2010.
- [13] Rahmadwati, G. Naghdy, M. Ros, C. Todd, and E. Norachmawati, "Classification Cervical Cancer Using Histology Images," *Proceeding 2010 Second International Conference on Computer Engineering and Application*, vol. 2, 2010.
- [14] Rahmadwati, G. Naghdy, M. Ros, C. Todd, and E. Norachmawati, "Cervical Cancer classification using Gabor filters," *1st IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2011*, pp. 48-52, 2011.
- [15] Yi-Wei and W. Jung-Hua, "Image segmentation based on region growing and edge detection," in *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, 1999, pp. 798-803 vol.6.
- [16] C. G. Loukas, G. D. Wilson, B. Vojnovic, and A. Linney, "An Image Analysis-Based Approach for Automated Counting of Cancer Cell Nuclei in Tissue Sections," *Cytometry* vol. Part A 55A, pp. 30-42, 2003.
- [17] A. Hafiane, F. Bunyak, and K. Palaniappan, "Fuzzy Clustering and Active Contours for Histopathology Image Segmentation and Nuclei Detection," *LNCS 5259*, 2008.
- [18] M. A Roula, A. Bouridane, and F. Kurugollu, "An Evolutionary Snake Algorithm for the Segmentation of Nuclei in Histopathological Images," *International Conference on Image Processing*, vol. 4, 2004.
- [19] Bamford P.C, "Segmentation of Cell Images with Application to Cervical Cancer Screening," PhD., Dept Computer Science and Electronical Engineering Brisbane, , University of Queensland., 1999.
- [20] C. K. Chow and T. Kaneko, "Automatic Boundary Detection of the Left Ventricle from Cineangiograms.," *Comput Biomed Res*, vol. 5, pp. 388-410, 1972.
- [21] He, Lei, Rodney Long , Samer Antani., George R.Thoma, "Local and Global Gaussian Mixture Modls for Hematoxylin and Eosin Stained histology Image Segmentation," *International Conference on Hybrid Intelligent System*, 2010.
- [22] Gonzales,Rafael and Richard E.Woods, *Digital Image Processing* third ed.: Prentice Hall, 2007.
- [23] He, Lei, Rodney Long , Samer Antani., George R.Thoma, "Histology image analysis for carcinoma detection and grading," *Computer Methods and Programs in Biomedicine*.
- [24] He,Lei, Rodney Long, Sameer Antani and George Thoma, "Computer-assisted diagnosis in cervical histopathology," *SPIE on Biomedical Optics & Medical Imaging*, 2010.
- [25] Price G.J, McCluggage, W. G., Morrison, M. L., McClean, G., Venkatraman, L., Diamond, J., Bharucha, H., Montironi, R., Bartels, P. H., Thompson, D., & Hamilton, P. W., "Computerized diagnostic decision support system for the classification of preinvasive cervical squamous lesions.," *Human Pathology*, vol. 34, pp. 1193-1203, 2003.
- [26] Arif,M and N. Rajpoot, "Classification of potential nuclei in prostate histology images using shape manifold learning," in *Machine Vision, 2007. ICMV 2007. International Conference on*, 2007, pp. 113-118.
- [27] B. Luck, Bovik, A. Rebecca, R , and Kortum.R, "Segmenting Cervical Epithelial Nuclei From Confocal Images Using Gaussian Markov Random Fields," *IEEE*, 2003.
- [28] Z. Li and K. Najarian, "Biomedical Image Segmentation Based On Shape Stability," *ICIP*, 2007.
- [29] Ilea, Dana Elena and Paul F.Whelan, "Color Image Segmentation Using a Spatial K-Means Clustering Algorithm," *Vision System Group*, 2009.
- [30] Y. Boykov, and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 2001, pp. 105-112 vol.1.
- [31] Seo,Natashi, "Normalized Cuts and Image Segmentation," *ENEE 731 Project* 2006.

- [32] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, 2000
- [33] C. Demir and B. Yener, "Automated Cancer Diagnosis based on Histopathological Images: a Systematic Survey," *Technical Report*, vol. 05-09.
- [34] EKhvedchenya, Brief Tutorial/Intro to the Mathematical Morphology in Image Processing. 2011
- [35] Belc ,Dan.I, "A Hybrid Wavelet Filter For Medical Image Compression," Doctor of Philosophy, Electrical and Computer Engineering, FAMU/FSU College Engineering, Florida, 2006.
- [36] H. Frederich Nauth, *Gynecological Cytology*, First ed. Germany: Thieme, 2007.
- [37] K.-B. Kim, D. Song, Y. Woo, A. An, J. Stefanowski, S. Ramanna, C. Butz, W. Pedrycz, and G. Wang, "Nucleus Segmentation and Recognition of Uterine Cervical Pap-Smears." vol. 4482: Springer Berlin / Heidelberg, 2007, pp. 153-160.
- [38] Koss.L and Myron R.M, *Koss' Diagnostic Cytology And Its Histopathologic Bases*, 5 ed. vol. 1: Lippincott williams & wilkin.
- [39] Reinhold, Wolfgang Schlote, "Methods for Shape Analysis of two-dimensional closed Contours - A biologically important, but widely neglected Field in Histopathology," *Electronic Journal of Pathology and Histology*, vol. 8, 2002.
- [40] Anynomous. CallC Cell counting. *User Manual*.
- [41] Shen, Liang, Rangaraj M. Rangayyan, and J. E. Leo Desautels, "Application of Shape Analysis to Mammographic Calcifications," *IEEE Transaction on Medical Imaging*, 1994.
- [42] Mahanta L.B, Dilip Ch. Nath, 1 Chandan Kr. Nath, "Cervix Cancer Diagnosis from Pap Smear Images Using Structure Based Segmentation and Shape Analysis," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3, 2012.
- [43] James H. Nelson, MD, Hervy E. Averette, MD, Ralph M. Richart, MD, "Cervical IntraepithelialNeoplasia (Dysplasiaand Carcinoma in Situ) And Early InvasiveCervical Carcinoma," *CA: A Cancer Journal for Clinicians*, vol. 39, 1989.