SEMI-AUTOMATIC FEATURES EXTRACTION OF CERVICAL CELLS

Oleh

VEERAYEN A/L MOHANADAS

Disertasi ini dikemukakan kepada
UNIVERSITI SAINS MALAYSIA

Sebagai memenuhi sebahagian daripada syarat keperluan
untuk ijazah dengan kepujian

SARJANA MUDA KEJURUTERAAN (KEJURUTERAAN ELEKTRONIK)

Pusat Pengajian Kejuruteraan
Elektrik dan Elektronik
Universiti Sains Malaysia

MAC 2005

ABSTRACT

This project is entitled ‘Semi-automatic Features Extraction of Cervical Cells’. The project is aimed to create a user friendly software which can be able to analyze Pap smear images via image processing. Cytological screening using the Pap smear test is the most effective strategy for the detection of precancerous state and consequent control of cervical cancer. Cytological samples that are taken from Pap smear test will undergo further analysis to detect the degree of abnormality of the cervical cells. The results of the abnormality of the samples can be inaccurate since some types of the medical images are blurring and highly affected by unwanted noise. Those bottlenecks in the medical images are believed that can be reduced via implementations of an *adaptive fuzzy c-means* (AFCM) and *moving k-means* (MKM) clustering techniques. These clustering techniques were used to segment the Pap smear images and later the features of the cells were extracted using *region growing based feature extraction* (RGBFE) technique. The performance of AFCM and MKM were analyzed based on the segmentation results of 6 Pap smear images. In overall, MKM was produced much better images than AFCM. Although the results have revealed that AFCM was suffering from centre redundancy and poor final centres in most of the cases, but it has also shown an advantage over MKM where AFCM was not sensitive to initial centres.

ABSTRAK

Projek yang dijalankan adalah bertajuk ‘Pengekstrakan Ciri-Ciri Sel Barah daripada Pangkal Rahim Secara Semi-automasi’. Projek ini adalah bertujuan untuk menghasilkan perisian yang mesra pengguna yang boleh memproses imej digital palitan Pap. Proses diagnosis secara “cytological” melalui ujian palitan Pap merupakan kaedah yang paling berkesan dalam pengesanan awal barah pangkal rahim. Sampel ujian palitan Pap akan melalui proses seterusnya di mana ia akan dianalisis supaya sel-sel yang tidak normal dapat dikesan pada peringkat awal lagi. Proses pengesanan sel-sel yang tidak normal boleh menjadi lebih sukar sekiranya imej palitan Pap tersebut kabur dan mempunyai kesan hingar yang tinggi. Kekurangan dalam imej palitan Pap tersebut dipercayai dapat dikurangkan melalui dua kaedah pengelompokan iaitu *adaptive fuzzy c-means* (AFCM) dan *moving k-means* (MKM). Kedua-dua jenis kaedah pengelompokan tersebut telah digunakan dalam sistem yang dibina untuk meruas imej digital palitan Pap. Imej digital palitan Pap yang telah diruas akan diekstrak ciri-cirinya dengan menggunakan teknik pengekstrakan iaitu *region growing based feature extraction* (RGBFE). Kedua-dua teknik peruasan telah diuji ke atas 6 imej digital palitan Pap. Daripada keputusan yang diperolehi, didapati MKM telah menunjukkan kebolehan peruasan imej digital palitan Pap yang tinggi berbanding AFCM. AFCM telah menghadapi masalah pertindihan pusat bersama masalah pusat akhir yang kurang baik dalam kebanyakan keadaan. Namun begitu, AFCM telah memaparkan satu kelebihan berbanding MKM di mana ia tidak sensitif kepada nilai pusat awal.

CONTENTS

ACKNOWLEDGEMENT

This project would not have been a success without the help and attention of many parties. I would like to take this opportunity to express my sincere gratitude to all those who have contributed in completing this project.

Firstly, I would like to express my appreciation to my final year project supervisor, Pn. Dzati Athiar Ramli as well as my co-supervisor, Dr. Nor Ashidi Mat Isa for their expert supervision and constant attention throughout my final year project. It has been a privilege to work under such fine supervisory. Their comments enabled me to produce a better quality project and the final report.

Furthermore, I would like to thank Dr. Mohd Yusof Mashor who has spent valuable time for me to clear all my doubts and misunderstanding in clustering algorithms. That too led me to complete this project successfully. And not forgotten Miss. Nazahah Mustafa who gave me some beneficial guidelines regarding image processing.

Finally, I wish to thank my parents and my friends who had supported and motivated me throughout the completion of this project.

Thank you.

CHAPTER 1

INTRODUCTION

1.1 Introduction

As the world is moving towards the glory of the information technology systems, many conventional methods are being replaced by faster and sufficient methods which use the latest and well developed technologies. Such a transformation can be seen in medical imaging as well. Image processing techniques have been seen as the most potential techniques in medical imaging to overcome the bottlenecks of the conventional methods that were used in diagnosing processes.

Pap smear images need consideration since in developing countries, cervical cancer is the leading cause of death from cancer. Unlike other cancers that cause pain, noticeable lumps or other early symptoms, cervical cancer has no telltale symptoms until it is so advanced that usually unresponsive to treatment (WebMD, 2002). However, most cervical cancer takes many years to develop from normal to dangerous stage. The death rate related to cervical cancer can be substantially reduced through early detection and treatment. Hence, Pap smear images need to be of high quality so that diagnosing processes are carried out more efficiently.

1.2 Objective of the Project

This project has been carried out to design a reliable system that would be able to extract the features of cervical cells by using two clustering techniques; an *adaptive fuzzy c-means* (AFCM) and *moving k-means* (MKM). These two algorithms were tested on Pap smear image samples that are in bitmap format.

This project was driven by the need to overcome some deficiencies in the quality of the Pap smear image samples. The Pap smear image samples are blurred and highly affected by unwanted noises, such as blood, air artifact, vagina discharge, and etc.

AFCM and MKM clustering techniques have been chosen as both the techniques are capable in reducing three main clustering problems; dead centres, local minima and centre redundancy. These two clustering techniques are also capable in producing good quality of segmentation image. But, theoretically, MKM clustering technique has proven that it is reliable to be used in Pap smear image processing.

Meanwhile, AFCM clustering algorithm is an alternative or improvement to the standard *fuzzy c-means* clustering algorithm (Bezdek *et al*. 1987). *Fuzzy c-means* clustering algorithm has been facing problems of poor local minima and initial centre sensitivity. Thus, Mashor (2001) has designed AFCM clustering algorithm to reduce those problems in *fuzzy c-means* clustering algorithm. Since the algorithm is based on an adaptive technique, the algorithm can be implemented using off-line or on-line techniques. In addition to that, the performance of AFCM in *radial basis function* (RBF) network has shown that it is very potential to be implemented in Pap smear image processing. Hence, its reliability in the Pap smear image processing is tested through this project.

The system that has been developed has features which are user friendly. The user can do segmentation on the selected image sample by using either AFCM or MKM clustering techniques and can extract the features of the cell as well. The features that are mentioned here are size and grey level values of the important structures of the cell; nucleus and cytoplasm.

1.3 Scope of the Project

Before starting with the project, the concepts of the clustering, Pap smear test, and image processing were gained mainly via internet, journals, and thesis which was done by previous students. In addition to that, the basic concept and the flow of the algorithm were gained from the supervisor and other lecturers as well. These things will be very important in completing this project successfully.

Then, the system was built using Borland C++ Builder version 5.0 software. The developed system is based on user friendliness. The system can perform image segmentation on Pap smear images (in bitmap format only) either by using AFCM or MKM clustering technique. MKM was chosen to be implemented in this project because of its good performance in Pap smear image processing in the previous studies. AFCM has been seen to give good performance in RBF network. Its performance in Pap smear image processing was studied via this project. The developed system was also able to extract features of the segmented cells; size of the nucleus and cytoplasm as well as their grey level values.

Later, the developed system was tested on 6 Pap smear image samples. Both the segmentation techniques, AFCM and MKM, were compared in terms of quality of the segmented image. The results were analyzed and discussed further.

Last but not least, conclusions have been drawn based on the segmentation results. The advantages and disadvantages of the algorithms will also be discussed.

1.4 Guidelines of the Report

Chapter 1 is an introduction to the project. The subtopics that have been included in this chapter are objective of the project, scope of the project, and the summary of each chapter in this report.

Chapter 2 touches upon the literature review. The topics that covered in this chapter are cervical cancer and Pap smear test, basic concepts of digital images, clustering, and problems regarding clustering. The chapter ends with an explanation of software that was used in this project (Borland C++ Builder version 5.0).

Chapter 3 is all about Pap smear image processing and algorithms that have been implemented through this project. This chapter explains problems in Pap smear image processing and the way Pap smear image processing was carried on in this project. The chapter also includes all the algorithms that were used.

Chapter 4 discusses on the results of the segmentation on Pap smear image samples. There are 6 Pap smear image samples that have been tested. Earlier in this chapter is an explanation about the functionality of the developed system. The features of the extracted cells are also included in this chapter.

Chapter 5 is the last chapter which includes conclusions and recommendations.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Digital image processing technique is being improved as more ideas are being contributed by many intellectual persons in the corresponding field. There are several algorithms that have been implemented in medical imaging to segment an image such as adaptive k-means clustering algorithm (Chen, *et al*., 1998). But, in this project two clustering algorithms were implemented; an *adaptive fuzzy c-means* (AFCM) (Mashor 2001) and *moving k-means* (MKM) (Mashor, 2000). Both algorithms were tested on Pap smear images which were in bitmap format.

2.2 Cervical Cancer

Cervical cancer is the most common cancer after cancer of breast, in women worldwide with an annual incidence estimated in excess of 440,000 cases. However, in Malaysia it comes after ovarian cancer. No other cancer better documents the remarkable effects of prevention, early diagnosis and treatment on the mortality rate than cervical cancer.

The cervix is an area of a woman's body that can develop abnormal cells and can, in some women, develop cancer. Abnormal cells are called dysplasia or "precancerous." Cervical cancer occurs when abnormal cervical cells are taken over by the malignant cells. Cervical cancer can spread to other female organs, such as the uterus, vagina, and ovaries. It can then spread to nearby organs such as the bladder, colon, and lymph nodes.

2.2.1 **Causes of Cervical Cancer**

The exact cause of cervical cancer is not known, but certain things appear to increase the risk:

- Human papilloma virus (HPV)

  Specific types of the human papilloma virus (the same virus that causes genital warts) are linked with 95% of cases of cervical cancer. HPV is passed on through sex and usually causes no symptoms at all.

- Sex behavior

  Starting to have sex at an early age may expose the cervix to HPV at an especially susceptible time. Plus, the more sexual partners a woman has, the greater the risk of getting HPV.

- The pill

  The contraceptive pill may increase the risk of cervical cancer because barrier methods such as condoms, which give some protection from HPV, are less likely to be used.

- Unhealthy lifestyle

  Women who smoke are more likely to be affected than non-smokers, and, as with most cancers, its thought that diet can affect the risk. A healthy diet is recommended, including fruit, vegetables, fiber-rich and starchy foods.

- Immune deficiency

  A weakened immune system increases the risk. Causes of immune deficiency include autoimmune diseases, such as rheumatoid arthritis, and human immunodeficiency virus (HIV) infection.

- History of abnormal cells

  If abnormal cells (dyskaryosis) have previously been found on the cervix, women are at a higher risk. However, many women who have cervical cancer do not appear to have any of these risk factors.

2.2.2 **Pap Smear Test**

Prevention, early diagnosis and treatment have been shown to reduce mortality due to cervical cancer. Cytological screening using the Papanicolou (Pap) smear test remains the most effective strategy for the detection of precancerous state and consequent control of cervical cancer.

During a Pap test, a small sample of cells from the surface of the cervix is collected by a health professional. The sample is then spread on a slide (Pap smear) or mixed in a liquid fixative and sent to a lab for examination under a microscope. The cells are examined for abnormalities that may indicate cancer or changes that could lead to cancer. Any abnormal cells are classified according to their degree of abnormality.

Classification of cervical cancer has been carried out in a variety of ways. The new and commonly used is the Bethesda system. Abnormal cervical cells are classified into two types; low grade intraepithelial lesions (LSIL) and high grade intraepithelial lesions (HSIL). Cytopathologists differentiate both types of abnormal cervical cells and normal cells based on several morphologies.

The abnormal cervical cells show changes in nucleocytoplasmic ratio. The cytoplasm size decreases but the nucleus size increases from normal cells to HSIL cells through LSIL cells (Crum, 1994). This phenomena increase the nucleus-to-cytoplasm ratio. Besides that, the abnormal cervical cells also show changes in colour (grey level) of nucleus and cytoplasm (WebMD, 2002). The grey levels for the cells' structures become darker from normal cells to HSIL cells through LSIL cells.

2.3 **Digital Image**

An image is a picture, photograph, display or other form which gives a visual representation of an object or scene. In terms of digital image processing, an image, $a[m,n]$ described in a 2D discrete space is derived from an analog image $a(x,y)$ in a 2D continuous space through a *sampling* process that is referred to as digitization.

2.3.1 **Basic Concepts**

The 2D continuous image $a(x,y)$ is divided into N *rows* and M *columns*. The intersection of a row and a column is termed as a *pixel*. The digital image can be conveniently represented by an N x M matrix **A** of the form:

$$\mathbf{A} = \begin{bmatrix} A(0,0) & A(0,1) & \dots & A(0,M-1) \\ A(1,0) & A(1,1) & \dots & A(1,M-1) \\ \vdots & \vdots & & \vdots \\ A(N-1,0) & A(N-1,1) & \dots & A(N-1,M-1) \end{bmatrix}$$

Figure 2.1: Image representation in form of matrix

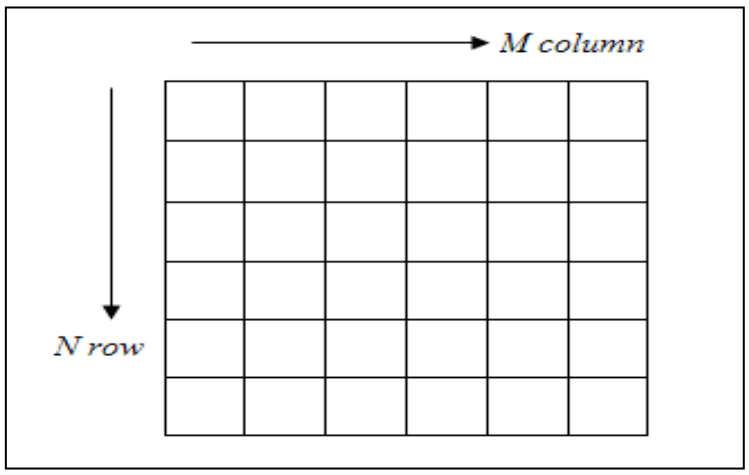

Figure 2.2: Image representation in form of grid

The value assigned to the integer coordinates [*m*,*n*] with {*m*=0,1,2,...,*M*-1} and {*n*=0,1,2,...,*N*-1} is *a*[*m*,*n*]. The coordinates [*m*,*n*] have the values in the range [0,…,255] for 8 bit images. Therefore, they can be represented as *characters* in the C++ language. In most cases *a*(*x*,*y*) is actually a function of many variables including depth (*z*), color (λ), and time (*t*).

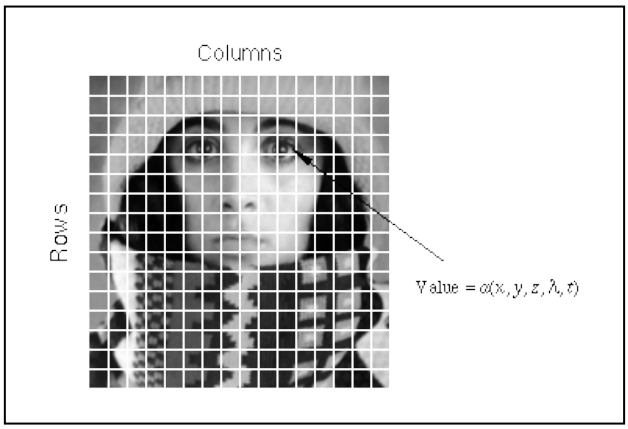

Figure 2.3: Digitization of a continuous image. The pixel at coordinates [*m*=10, *n*=3] has the integer brightness value 110.

The image shown in Figure 1 has been divided into *N* = 16 rows and *M* = 16 columns. The value assigned to every pixel is the average brightness in the pixel rounded to the nearest integer value. The process of representing the amplitude of the 2D signal at a given coordinate as an integer value with *L* different gray levels is usually referred to as amplitude quantization or simply *quantization*.

2.4 **Clustering**

Clustering is a classification technique. Given a vector of *N* measurements describing each pixel or group of pixels (i.e., region) in an image, a similarity of the measurement vectors and therefore their clustering in the *N*-dimensional measurement space implies similarity of the corresponding pixels or pixel groups. Similarity between image regions or pixels implies clustering (small separation distances) in the feature space. Clustering methods were some of the earliest data segmentation techniques to be developed. AFCM and MKM are two types of clustering techniques in image segmentation.

The main concept behind the clustering technique is the minimization of the total distance between the data and the centres so that the centres can properly represent the data. When the centres are updated during the clustering technique there is the possibility to centre redundancy to be happen. This centre redundancy causes misrepresentation of data at the end of segmentation. It is always a very good idea to look after this problem so that none of the data (pixels) missed during the segmentation technique.

2.4.1 **Clustering Problems**

Most clustering algorithms work on the assumption that the initial centres are provided. The search for the final clusters or centres starts from these initial centres. Without a proper initialization, such algorithms may generate a set of poor final centres and this problem can become serious if the data are clustered using an on-line

clustering algorithm. In general, there are three basic problems that normally arise during clustering:

- Dead centres
- Local minima
- Centre redundancy

Dead centres are centres that have no members or associated data. Dead centres are normally located between two active centres or outside the data range. This problem may arise due to bad initial centres, possibly because the centres have been initialized too far away from the data. Therefore, it is a good idea to select the initial centres randomly from the data or to set the initial centres to some random values within the data range. However, this does not guarantee that all the centres are equally active (i.e. have the same number of members). Some centres may have too many members and be frequently updated during the clustering process whereas some other centres may have only a few members and are hardly ever updated.

The centres should be selected to minimize the total distance between the data and the centres so that the centres can properly represent the data. During the clustering process, the centres are adjusted according to a certain set of rules such that the total distance is minimized. However, in the process of searching for the global minima the centres can be frequently become trapped at local minima. Poor local minima may be avoided by using algorithms such as simulated annealing, stochastic gradient descent and genetic algorithms.

The centres should be sufficient enough to represent the identified data. However, as the number of centres increases, the tendency for the centres to be located at the same position or very close to each other is also increased. There is no point in adding extra centres if the additional centres are located very close to the centres that already exist.

2.5 **Borland C++ Builder version 5.0 Software**

Borland C++ Builder version 5.0 was used in this project to develop the software of image processing of the Pap smear image. C++ Builder is a new *rapid application development* (RAD) product for writing C++ applications. With C++ Builder we can write C++ Windows programs more quickly and more easily. We can create Win32 console applications or Win32 GUI (graphical user interface) programs.

The C++ Builder IDE (which stands for *integrated development environment*) is divided into three parts. The top window might be considered the main window. It contains the Toolbar on the left and the Component Palette on the right. The Toolbar gives one click access to tasks like opening, saving and compiling projects. The Component Palette contains a wide array of components that can be dropped onto the forms. (*Components* are things like text labels, edit controls, list boxes, buttons, and so on.)

Below the Toolbar and Component Palette and glued to the left side of the screen is the Object Inspector. Through the Object Inspector component's properties and events can be modified. A component's *properties* control how the component operates. For example, changing the Color property of a component will change the

background color of that component. Events occur as the user interacts with a component. For example, when a component is clicked, an event fires and tells Windows that the component was clicked. The user can write code that responds to those events, performing specific actions when an event occurs.

To the right of the Object Inspector is the C++ Builder workspace. The workspace initially displays the Form Editor. In C++ Builder a form represents a window in the user's program. The user uses the Form Editor to place, move, and size components as part of the form creation process. Then, the Code Editor is where the user types code when writing programs. The Object Inspector, Form Editor, Code Editor, and Component Palette work interactively as the user builds applications.

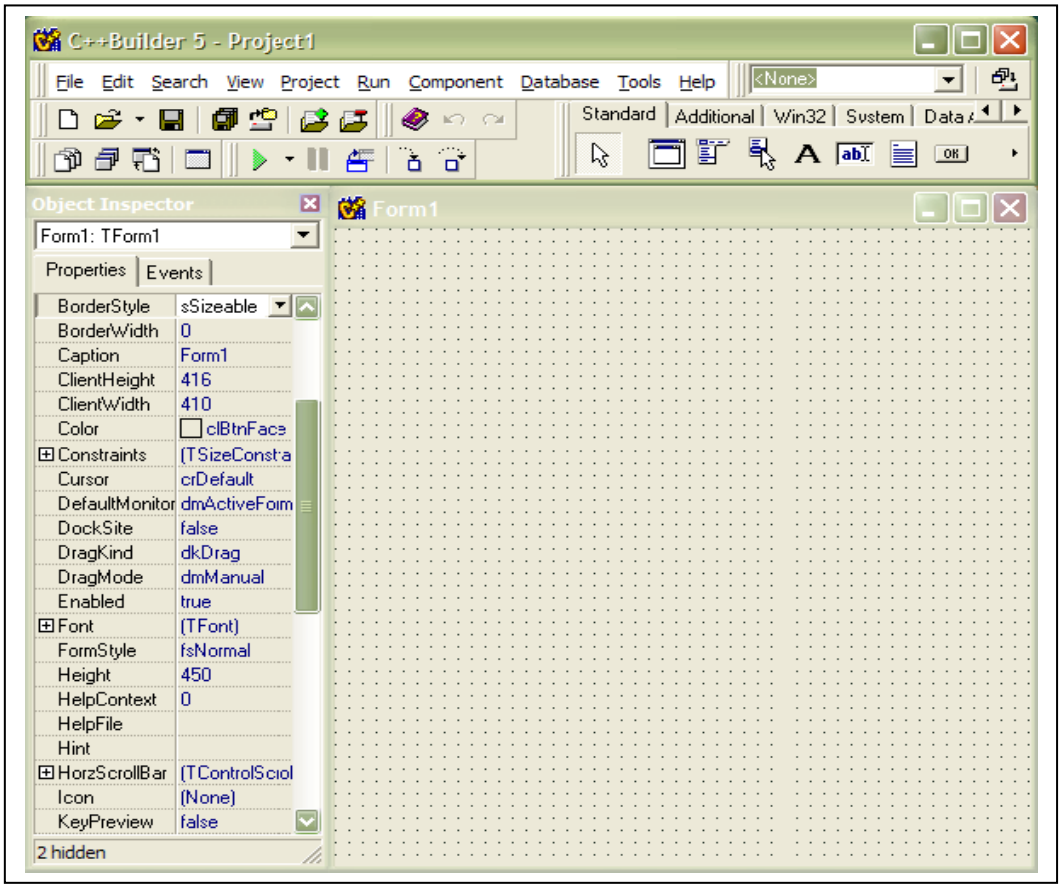

Figure 2.4: The C++ Builder IDE and the initial blank form

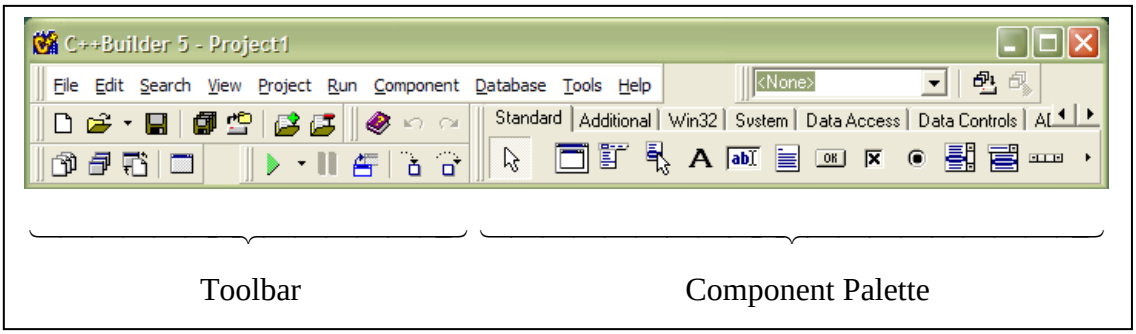

Figure 2.5: Speedbar and Component Palette

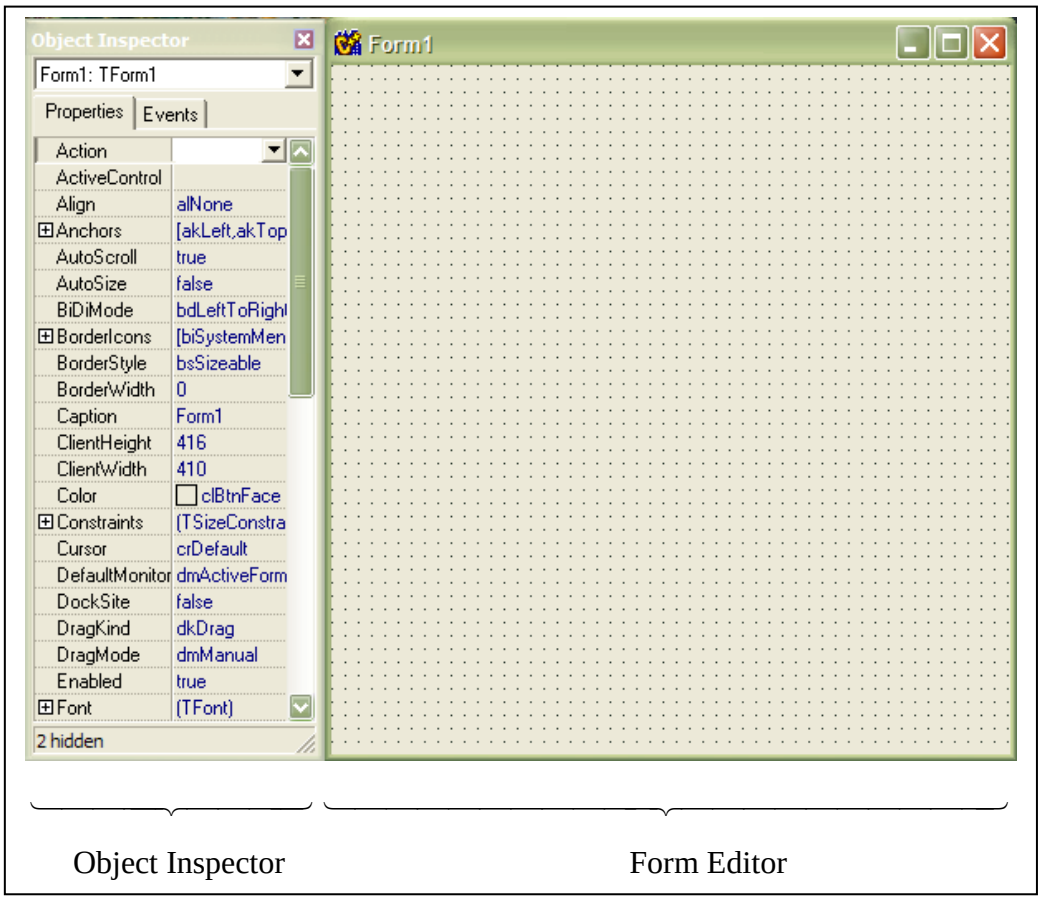

Figure 2.6: Object Inspector and Form Editor

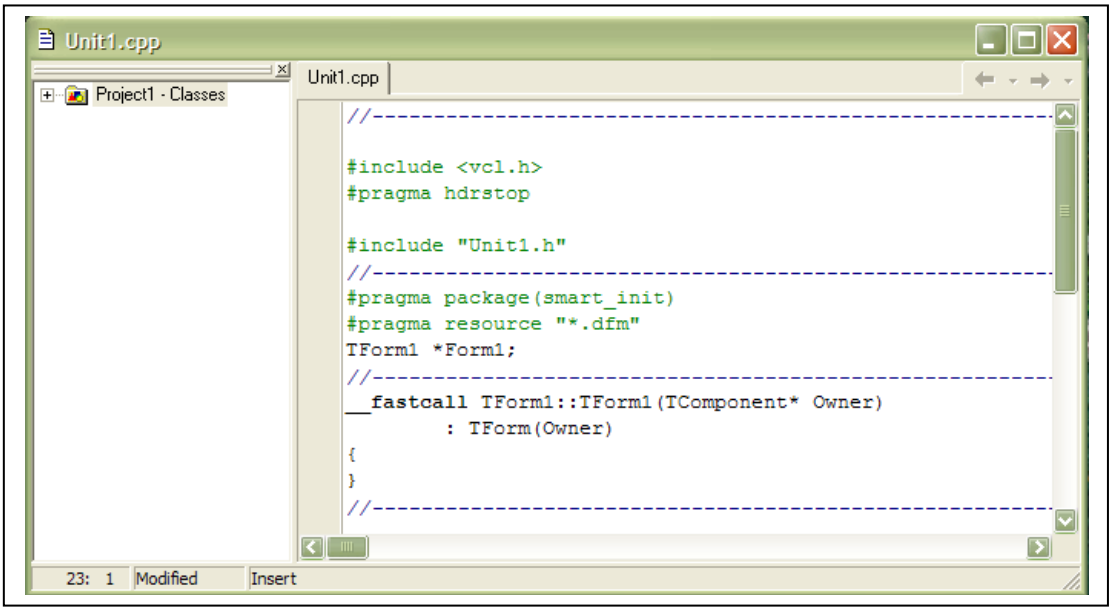

Figure 2.7: Code Editor

When the user first creates the project, C++ Builder creates a minimum of six files:

- The project source file
- The main form source file
- The main form header file
- The main form resource file
- The application resource file
- The project makefile

The *project source file* is the file that contains the WinMain ( ) function and the other C++ Builder startup code. The *main form source file* and *main form header file* are files that contain the class declaration and definition for the main form's class. The *main form resource file* and *application resource file* are binary files that describe the main form and the application's icon. The *makefile* is a text file that contains

information about the compiler options that the user has set, the names of the source files and forms that make up the project, and what library files have to be included.

There are several types of files used in C++ Builder. But, the minimum set of files consists of the **.**CPP, **.**DFM, **.**H, and **.**MAK files. All other files are files that C++ Builder will re-create when the user compiles the program. Those four files' extension and the description are showed in Table 2.1.

Table 2.1: C++ Builder extension files and its description

Extension	Description
.CPP	The C++ source files. There will usually be one for each unit and one for the main project file, as well as any other source files that the user adds to the project.
.DFM	The form file. This file is actually a binary resource file (**.**RES) in disguise. It is a description of the form and all its components. Each form has its own **.**DFM file.
.H	C++ header files that contain class declarations. These could be C++ Builder-generated files or the user's own class headers.
.MAK	The project makefile. This is a text file that contains a description of which files C++ Builder needs to compile and link.

CHAPTER 3

PAP SMEAR IMAGE PROCESSING

3.1 Introduction

Even though Pap test is very reliable for detecting early abnormal cell changes that could lead to cervical cancer, the determination of that abnormal cells can be missed due to technical or human errors. Studies by Othman *et al.* (1997) and Hislop *et al.* (1994) have shown that some Pap smear cytology images are blurred and highly affected by unwanted noises, such as blood, air artifact, vagina discharge etc. These problems can hide and obscure the important morphologies of cervical cells. Thus, the Pap smear cytology image will be referred as inadequate samples for cervical cancer screening process.

3.2 Pap Smear Image Processing

The problem faced in Pap test as elaborated in introduction has driven to continuous researches on Pap smear image to produce a better image that could ease the diagnosing process in detecting abnormal cervical cells. Segmentation technique such as *moving k-means* has proved that it is very useful to segment the Pap smear image into three important structure of the cell; nucleus, cytoplasm and background.

Clustering technique is one of the earliest techniques used to do segmentation on digital images. In this project, an *adaptive fuzzy c-means* (AFCM) and *moving k-means* (MKM) clustering techniques have been used to segment the Pap smear image samples. The Pap smear image samples were in bitmap format. Pap smear images were segmented into three regions through segmentation technique. These regions are

nucleus, cytoplasm and background which being represented as black (grey level of 0), gray (grey level of 127), and white (grey level of 255) respectively. Later, size of the nucleus and cytoplasm as well as their grey level values were extracted using *region growing based features extraction* (RGBFE) technique.

3.3 Algorithm

Even though there is several numbers of segmentation techniques that have been designed such as *k-means* and *fuzzy c-means*, AFCM and MKM clustering techniques have been used to segment the image of cervical cells through this project. Both the techniques were compared based on the quality of image that was produced after segmentation. Later, the RGBFE technique was used to extract the size and grey level of the both nucleus and cytoplasm of the cell.

3.3.1 Adaptive Fuzzy *c*-means Clustering Algorithm

AFCM clustering algorithm is proposed by Mashor, (2001) as an alternative to the *fuzzy c-means* clustering algorithm. AFCM was designed to reduce the three clustering problems that have been discussed in Chapter 2.4.1. The algorithm updates all the centre locations after each data sample is presented hence the centres are updated more frequently.

The AFCM clustering algorithm has been implemented as:

1) Initialize the centres and the adaptation constants $c_j(0)$, $\mu(0)$ and q.

2) Calculate the Euclidean distances $d_j(t)$ between the data sample $v(t)$ and the centre $c_j(t)$ and find the shortest distance $d_s(t)$ and the longest distance $d_l(t)$; together with the associated centres, $c_s(t)$ and $c_l(t)$.

3) For $j = 1$ to n_c,

   a) Update the mean square distance between the centres and the data $v(t)$

$$\gamma(t) = \frac{1}{n_c}\sum_{k=1}^{n_c}[\| v(t) - c_k(t) \|]^2, \tag{3.1}$$

   b) If $j \neq s$, update the centre according to

$$\Delta c_j(t) = \mu(t)\vartheta(t)[v(t) - c_j(t-1)], \tag{3.2}$$

   where

$$\vartheta(t) = \begin{cases} D_l(t)D_j(t)\exp[-D_a(t)] \\ \qquad\qquad if\ \ d_j(t) > 0, \\ D_l(t)\exp[-D_a(t)] \\ \qquad\qquad if\ \ d_j(t) = 0, \end{cases}$$

   $D_l(t) = \gamma(t)/d_l^2(t)$, $D_j(t) = d_a^2(t)/d_j^2(t)$ and $D_a(t) = d_a^2(t)/\gamma(t)$, with $a = s$ if $d_s(t) > 0$ and $a = z$ is $d_s(t) = 0$ ($d_z(t)$ is the smallest non-zero distance between c_j and $v(t)$) and

   c) Update $c_s(t)$ according to

$$\Delta c_s(t) = \mu(t)\varphi(t)[v(t) - c_s(t-1)], \tag{3.3}$$

where

$$\varphi(t) = \begin{cases} \exp\left(-\frac{d_s^2(t)}{\gamma(t)} \right) & if \ d_s(t) > 0, \\ 0 & if \ d_s(t) = 0. \end{cases} \tag{3.4}$$

4) Check the distances between $c_s(t)$ and the rest of the centres, $h_k(t) = \| c_s(t) - c_k(t) \|$, $k = 1,2,...,n_c$ and $k \neq s$, and, if the shortest distance $h_c(t) < \mu(t) d_a(t)$, move the nearest centre $c_c(t)$ to a new location according to

$$\Delta c_c(t) = -\frac{\mu(t) d_a^2(t)}{d_l^2(t)} [c_c(t) - c_s(t)]. \tag{3.5}$$

5) $t = t + 1$ and repeat steps 2-4 for each sample of data.

The adaptation rate $\mu(t)$ for on-line implementation is updated according to

$$\mu(t) = \mu(0) \exp\left(-\frac{qt^2}{(n_c)^2} \right) + \frac{\exp[-q\mu(t-1)]}{n_c}, \tag{3.6}$$

and for off-line implementation the adaptation rate is given as

$$\mu(t) = \mu(0) \exp\left(-\frac{qt^2}{n_c^3} \right) + \frac{\exp[-q\mu(t-1)]}{n_c^{1/2}} \tag{3.7}$$

where $0 \leq \mu(0) \leq 1.0$, q is a constant $(0 \leq q \leq 1.0)$ and N is the number of training data. In general, q should increase with increasing number of centres and $\mu(0)$ is selected arbitrarily. The off-line adaptive fuzzy clustering normally has a smaller value of q than the on-line adaptive fuzzy clustering algorithm. A typical value of $\mu(0)$ is between 0.80 and 0.95 and q is between 0.01 and 0.30.

3.3.2 **Moving *k*-means Clustering Algorithm**

The algorithm proposed by Mashor, (2000) is based on non-adaptive clustering technique. The algorithm is called *moving k-means* (MKM) clustering because during the clustering process, the fitness of each centre is constantly checked and if the centre fails to satisfy a specified criterion the centre will be moved to the region that has the most active centre.

The MKM clustering algorithm has been implemented as:

1) Initialize the centres and α_0, and set $\alpha_a = \alpha_b = \alpha_0$.

2) Assign all data to the nearest centre and calculate the centre positions using equation:

$$c_j = \frac{1}{n_j} \sum_{i \in c_j} v_i \tag{3.8}$$

3) Check the fitness of each centre using equation:

$$f\left(c_j\right) = \sum_{i \in c_j} \left(\left\| v_i - c_j \right\| \right)^2; \qquad j = 1,2,\ldots,n_c; \qquad i = 1,2,\ldots,N \tag{3.9}$$

4) Find c_s and c_l, the centre that has the smallest and the largest value of $f(\cdot)$.

5) If $f\left(c_s\right) < \alpha_a \, f\left(c_l\right)$,

(5.1) Assign the members of c_l to c_s if $v_i < c_l$, where $i \in c_l$, and leave the rest of the members to c_l.

(5.2) Recalculate the positions of c_s and c_l according to:

$$\left.\begin{aligned} c_s &= \frac{1}{n_s} \sum_{i \in c_s} v_i \\ c_l &= \frac{1}{n_l} \sum_{i \in c_l} v_i \end{aligned}\right\} \tag{3.10}$$

Note: c_s will give up its members before step (5.1) and, n_s and n_l in equation (3.10) are the number of the new members of c_s and c_l respectively, after the reassigning process in step (5.1).

6) Update α_a according to $\alpha_a = \alpha_a - \alpha_a / n_c$ and repeat step (4) and (5) until $f(c_s) \geq \alpha_a \ f(c_l)$

7) Reassign all data to the nearest centre and recalculate the centre positions using equation (3.8):

8) Update α_a and α_b according to $\alpha_a = \alpha_0$ and $\alpha_b = \alpha_b - \alpha_b / n_c$ respectively, and repeat step (3) to (7) until $f(c_s) \geq \alpha_b \ f(c_l)$.

where α_0 is a small constant value, $0 < \alpha_0 < \frac{1}{3}$. The computational time will increase as the values of α_0 get larger. Hence α_0 should be selected to compromise between good performance and computational load. The centres for the algorithm can be initialised to any values but a slightly better result can be achieved if the centres are initialised within the input and output data range. If the centres are badly initialized, then α_0 should be selected a little bit bigger (typically > 0.2).

3.3.3 **Region Growing Based Features Extraction Algorithm**

Region growing technique has successfully been used as segmentation technique of digital images (Ooi *et al*., 2000). The potential use of thresholding the region growing technique as features extraction technique was utilized by Mat Isa *et al.* (2003). The proposed technique was called *region growing based features extraction* (RGBFE). But in this project, the features of the cervical cells were extracted after implementing either AFCM clustering technique or MKM clustering technique instead

of using the thresholding technique. As shown in Figure 3.1, RGBFE was used to extract the size and grey level of certain region of interest on a digital image.

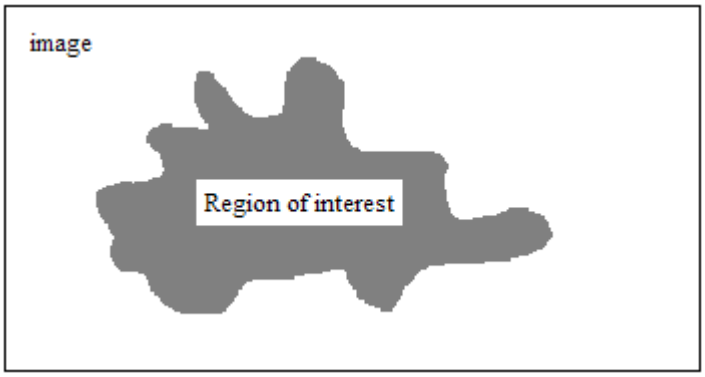

Figure 3.1: The region of interest for features extraction process

Referring to Figure 3.1, the size of the region is calculated as a total number of pixels in the region and given by the following equation.

$$Size = Total\ of\ pixels\ in\ the\ region \tag{3.11}$$

The grey level of the region is calculated as mean value of all pixels in the region and given by the following equation.

$$Grey\ level = \frac{Total\ of\ \ grey\ level\ \ for\ all\ \ pixels\ in\ the\ region}{Total\ of\ \ pixels\ in\ the\ region} \tag{3.12}$$

In the RGBFE technique, the user needs to determine the region of interest by clicking mouse on any pixels in the region. The algorithm has been implemented as:

1) Determine the threshold value.

2) Click mouse in the region of interest.

Note: The pixel, which mouse is clicked on it will be used as initial seed pixel.