Faculty of Cognitive Sciences and Human Development

FEATURE EXTRACTION AND CLASSIFICATION: A CASE STUDY OF CLASSIFYING A SIMULATED DIGITAL MAMMOGRAM IMAGES USING SELF-ORGANIZING MAPS (SOM)

Lau Leh Teen

Kota Samarahan
2007

BORANG PENGESAHAN STATUS TESIS **Gred:**

JUDUL : _____
_____
_____

SESI PENGAJIAN :_____

Saya _____
(HURUF BESAR)

mengaku membenarkan tesis * ini disimpan di Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Malaysia Sarawak
2. Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dibenarkan membuat salinan untuk tujuan pengajian sahaja
3. Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dibenarkan membuat pendigitan untuk membangunkan Pangkalan Data Kandungan Tempatan
4. Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi
5. ** sila tandakan (√)

☐	SULIT	(mengandungi maklumat yang berdarjah keselamatan atau kepentingan seperti termaktub di dalam AKTA RAHSIA RASMI 1972)
☐	TERHAD	(Mengandungi maklumat Terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
☐	TIDAK TERHAD	

_____ _____

(TANDATANGAN PENULIS) (TANDATANGAN PENYELIA)

Alamat Tetap:
_____
_____
_____

Tarikh : _____ Tarikh: _____

FEATURE EXTRACTION AND CLASSIFICATION: A CASE STUDY OF CLASSIFYING A SIMULATED DIGITAL MAMMOGRAM IMAGES USING SELF-ORGANIZING MAPS (SOM)

LAU LEH TEEN

This project is submitted in partial fulfillment of requirements for a Bachelor of Science with Honours Cognitive Science

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
2007

The project entitled Feature Extraction and Classification: A Case Study of Classifying a Simulated Digital Mammogram Images Using Self-Organizing Maps (SOM) was prepared by Lau Leh Teen and submitted to Faculty of Cognitive Sciences and Human Development in partial fulfillment of the requirements for a Bachelor of Science with Honours Cognitive Science.

Received for examination by:

(Dr. Teh Chee Siong)

Date:

Grade

ACKNOWLEDGEMENT

I would like to express my appreciation and gratitude to Mr. Teh Chee Siong for his guidance, comments and help in my effort to finish this project.

I would also like to give thank to my family members, friends and course mates for their help, support and advice throughout this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

ABSTRACT

FEATURE EXTRACTION AND CLASSIFICATION: A CASE STUDY OF CLASSIFYING A SIMULATED DIGITAL MAMMOGRAM IMAGES USING SELF-ORGANIZING MAPS (SOM)

Lau Leh Teen

Feature extraction is important in image processing and is a preliminary step to perform pattern classification. This project aims to propose a feature extraction technique. This feature extraction technique can be used to find five parameters which are the size, intensity, centroid X, centroid Y and region distribution of segmented regions. Several experiments have been conducted to verify the proposed algorithm and feature extraction results obtained will be used for the training of Neural Network classifier, Self-Organizing Maps (SOM). A set of training input data is used to train SOM. The accuracy of classification performance was acquired. A case study of breast cancer has been demonstrated in this study by using a simulated digital mammogram images. In this study, the results show that this system is able to perform the classification of mass with low intensity, mass with high intensity, cluster microcalcification, separate microcalcification and special case to detect abnormality of the digital mammogram images.

ABSTRAK

PENGEKSTRAKAN DAN PENGELASAN CIRI: SATU KAJIAN KES MENGENAI PENGELASAN SIMULASI IMEJ DIGITAL MAMMOGRAM MENGGUNAKAN SELF-ORGANIZING MAPS (SOM)

Lau Leh Teen

Pengekstrakan ciri adalah penting dalam pemprosesan imej dan merupakan langkah yang paling asas untuk melaksanakan pengelasan pola. Projek ini bertujuan untuk mencadangkan teknik pengekstrakan ciri. Pengekstrakan ciri ini boleh digunakan untuk mencari lima parameter iaitu saiz, kecerahan, lokasi pusat-X, lokasi pusat-Y dan taburan pembahagian kawasan. Beberapa eksperimen telah dijalankan untuk mengesahkan algoritma yang dicadangkan dan hasil pengelasan ciri yang diperolehi akan digunakan bagi melatih pengelasan rangkaian neural, Self-Organizing Maps (SOM). Satu set input data digunakan untuk melatih SOM. Ketepatan prestasi pengelasan akan diperolehi. Kajian kes mengenai kanser payudara telah ditunjukan dalam kajian ini menggunakan simulasi imej mammogram digital. Hasil kajian ini menunjukan sistem ini mampu melaksanakan pengelasan kumpulan mempunyai kecerahan rendah dan tinggi, kelompok pengapuran mikro, pengapuran mikro yang berasingan dan kes khas untuk mengenalpasti ketidaknormalan imej mammogram digital.

CHAPTER 1
INTRODUCTION

1.0 Introduction

The human perception has the capability to acquire, integrate, and interpret all visual information around us. It is a challenge to convey such capabilities to a machine with the aim of interpreting the visual information embedded in images, graphics and video. Thus, understanding the techniques of processing, segmentation, feature extraction, recognition and interpretation of such visual scenes are important (Acharya & Ray, 2005).

According to Wikipedia (2006), image processing is a form of information processing for images, such as photographs or frames of video case. Image processing techniques involve treating the image as a two-dimensional signal and applying standard signal processing techniques to it.

Besides that image processing is also concerned with taking one array of pixels as input and producing another array of pixels as output which in some way represents an improvement to the original array. This processing may remove noise in the images, improve the contrast of the images, remove blurring caused by movement of the camera during image acquisition, and correct the geometrical distortions caused by the lens (Marshall, 1994).

After that, the segmentation needs to perform so as to subdivide an image into a number of uniformly homogeneous regions. Segmentation is one of the most important elements in automate image analysis because at this step the objects or other entities of interest are extracted from an image for subsequent processing (Acharya & Ray, 2005).

A set of meaningful features such as size, texture, color and shape can be extracted after segment the image. This is because these features are important measurable entities which give measures of various properties of image segments (Acharya & Ray, 2005). Finally, each segmented object is classified to one of a set of meaningful classes based on the set of the extracted features.

There are a large number of applications of image processing in different spectrum of human activities. In this project, a case study of breast cancer will be demonstrated by using a simulated digital mammogram images.

Breast cancer is the most common type of cancer among women. According to Breast Health Information Centre (2005), there were 1,050,346 cases reported with 372,969 deaths from breast cancer world-wide in the year of 2000. The National Cancer Institute estimates that over a lifetime, there will be one out of eight women suffer from breast cancer.

In Malaysia, there were 3825 reported cases and 1707 died of breast cancer. Breast Health Information Centre (2005) claimed that the majority breast cancer patients are Chinese, followed by the Indians and Malays. Over a lifetime,

there will be one per nineteen chances a woman in Malaysia suffer from breast cancer.

However, early detection is a main factor for surviving and decreasing the risk of breast cancer deaths. MayoClinic (2006) stated that mammography plays a key role in early detection. Doctors can detect breast cancer nearly one to three years earlier through mammography before one might actually feel a lump in one's breast.

According to Sample (2003) common techniques from the field of image processing have been applied to digital mammograms in an effort to locate signs of cancer more precisely. The systems by drawing attention to areas of suspicion which are less noticeable and visible have the potential to greatly increase early detection.

1.1 Problem Statement and Motivation

According to Tian, Guo and Lyu (2005), feature extraction is a preliminary step before the classification because a classifier is able to recognize the interested object in image with extracted features.

Feature extraction is vital for subsequent image recognition and classification because unsuitable features can lead directly to wrong classification. Therefore, it is important to build a feature extraction system which can extract the optimal features from the segmented object (Yao, Jiang,Yi & Zhao, 2005).

Wang, Zhou and Geng (2005) also stated that classification is important because classification will find out the models that describe and distinguish the data classes and concepts. Based on the analysis of the training data set, this model is used to predict class to label the unknown object classes.

Medical imaging typically uses sensing methods with very different underlying physics such as magnetic resonance imaging (MRI), X rays and computer tomography (CT). It often deals with flexible, deformable and geometrically intricate objects. Thus, it is important to identify the location of abnormalities region for medical images in surgery, diagnosis and therapy evaluation (Kulkarni, 2001).

Kurkarni (2001) claimed that internal noise is inherent in the eye-brain system in very low-contrast situations. The radiologist will have difficulties in interpreting the images. Therefore, feature extraction and classification system are needed to identify the suspicious region in the image.

Furthermore, a large volume of medical images must be interpreted by radiologist daily. With a feature extraction system which can detect the suspicious region and extract the features, it will help the radiologists to identify the suspicious abnormalities region faster and correctly (Kulkarni, 2001).

1.2 Research Objective

1.2.1 General Objective

The general objective of this project is to design and develop a features extraction and classification system for a simulated digital mammogram images.

1.2.2 Specific Objectives

The specific objectives of this project are stated below:

a. To design and develop a feature extraction system to extract size, intensity, location and distribution of the segmented region in simulated digital mammogram images.

b. To classify the segmented region into five classes of abnormalities using selected Neural Network Classifier, Self-Organizing Maps (SOM).

1.3 Project Significant

a. A prototype that assists the radiologist to detect location of potential abnormalities precisely on the medical images.

b. This system will to reduce the human errors when interpreting the medical images by enhancing the accuracy of detecting abnormalities tissues in mammograms especially false negative.

1.4 Limitation of Project

There are several limitations in this project. First, the images used in this project are simulated digital mammogram images. This is because the pre-processing parts in this project are not advanced enough to process the real digital mammogram of breast cancer patients. This system can only process simulated prototype of digital mammogram images.

Besides that, the pre-processing part in this project is too simple. Improvement of the pre-processing of feature extraction and the classification system need to be performed in future. More features need to be extracted from the segmented image. This can assist the system to classify the image in the digital mammogram images precisely and accurately.

The last limitation of this project is that the data and information used in this project are not collected from the expert radiologists and doctors. Therefore, this system cannot classify the digital mammogram image for the real world problems.

CHAPTER 2
LITERATURE REVIEW

2.0 Introduction

This chapter discusses the fundamental steps in digital image processing and application of image processing. It is followed by the explanation of Artificial Neural Network (ANN) and application of ANN in the medical field. After that, Self-Organizing Map (SOM) algorithm and architecture will be discussed.

2.1 Definition of Image

Generally, an image is defined as a two-dimensional function, $f(x,y)$, where x and y are spatial coordinates and amplitude of f at any pair of coordinates. The amplitude of f is known as the intensity or gray level of the image at any particular point of coordinates (x,y). A digital image is composed of a finite number of elements and each of elements has a particular location and value. These elements are picture elements, image elements, pels and pixel. Pixel is the term most widely to represent the element of digital image (Gonzalez & Woods, 2001).

2.2 Digital Image Processing

Digital image processing refers to the transformation of an image to a digital format using digital computers. Both input and output of a digital image processing system are digital images (Pitas, 1993). Digital image processing is dealing with various images which consist of different complexity of data or information.

2.3 Fundamental Steps of Digital Image Processing

2.3.1 Image Acquisition

Image acquisition is the first stage in digital image processing. Image acquisition is an important step because if the image has not been acquired satisfactorily then intended tasks may not be achieved later. The image with its essential features will be converted into a set of digitized data to be processed by the system. The image acquisition consists of four phases which is illumination, image formation, image projection and image digitization (Galbiati, 1990).

There are a lot of file formats in which one may store the images in files and retrieve them from files. According to Acharya and Ray (2005), the most popularly used image file format standards are Tagged Image Format (.tif, .tiff), Portable Network Graphics (.png), JPEG (.jpg), MPEG (.jpg), MPEG (.mpg), Graphics Interchange Format (.gif), RGB (.rgb), RAS (.ras), postscript (.ps, .eps, .epsf), Portable Image File Formats, PPM, PGM, and PBM.

2.3.2 Image Enhancement

The objective of image enhancement is to process an image so that the result is more appropriate than the original image for a specific application. Enhancement approaches fall into two broad categories which are spatial domain methods and frequency domain methods. Spatial domain refers to image plane

itself and is based on direct manipulation of pixels in an image, while frequency domain methods are based on modifying the Fourier transform of an image (Gonzalez & Woods, 2001).

Enhancement technique can be employed in an attempt to improve the image in a subjective way. This non-specific operation will improve the visual appearance of image and the perceived quality of images suffering from either of these faults. Enhancements are usually used to improve the delectability of certain features in an image, so that the human can perceive it more easily (Lewis, 1990).

The spatial image enhancement techniques used for noise reduction or smoothing are spatial low pass, high-pass and band-pass filtering, unsharp masking and crisping, directional smoothing and median filtering.

2.3.3 Image Restoration

Restoration techniques can be employed for correction when degradations are present in an input image (Lewis, 1990). Restoration tries to reconstruct or recover an input image that has been degraded by using a prior knowledge of the degradation phenomenon. Thus, restoration techniques are oriented to model the degradation and apply the inverse process to recover the original image (Gonzalez & Woods, 2001).

2.3.4 Image Segmentation

Segmentation is a word used to describe a grouping process in which the components of a group are similar with respect to some feature or set of features (Lewis, 1990). Therefore, image segmentation is the first step in image analysis for extracting information from an image.

Segmentation divides an image into constituent regions or objects. The level to which the subdivision is carried depends on the problem being solved.

Segmentation should stop when the objects of interest in an application have been isolated because there is no point in carrying segmentation past the level of details required (Gonzalez & Woods, 2001).

Segmentation algorithm for monochrome images generally are based on one of two basic properties of intensity values which is discontinuity and similarity. The discontinuity is to partition an input image based on abrupt changes in grey level, while similarity is based on thresholding, region growing, region splitting and merging (Gonzalez & Woods, 1992).

2.3.5 Image Representation and Description

After image segmentation, a representation is chosen for transforming data into a form suitable for subsequence processing. Description or feature selection needed to highlight the features of interest. It deals with extracting features that result in some quantitative information of interest or features that are basic for differentiating one class of objects from another (Gonzalez & Woods, 1992).

2.3.6 Image Recognition and Interpretation

Recognition is the process that assigns a label to an object based on information provided by descriptors while interpretation involves assigning meaning to an ensemble of recognized images (Gonzalez & Woods, 1992). The approaches to image recognition are decision-theoretic and structural. Decision-theoretic recognition is based on representing patterns in vector form and then seeking approaches for grouping and assigning pattern vector into different classes. Structural method achieves pattern recognition by capitalizing precisely on structural relationship inherent in a pattern shape. Pattern is represented in symbolic form (Gonzalez & Woods, 2001).

2.4 Application of Digital Image Processing in Medical

The objective of most image processing applications is to extract meaningful information from the input image. In the medical field, image processing application is used to derive information about a patient's state of health from X-ray, ultrasound, magnetic resonance and others relevant image (Lewis, 1990). Digital image processing can enhances the contrast or code the intensity levels into color for easier interpretation of X-rays and other biomedical images (Gonzalez & Woods, 1992).

Image processing does not change the diagnosis of problem but it is claimed that doctors have been able to diagnose quicker from the processed image than from the original (Low, 1991).

Gradient magnitude based on region growing used to segment tumor from F-18-FDG PET images. This can reduce the time for manual post-processing due to over segmentation result from ordinary region growing based on intensity value (Yu et al, 2005).

2.5 Artificial Neural Network

'Artificial neural network' or ANN is the term used to describe a computer model assumption of the biological brain. It consists of a set of interconnected simple processing units which combine to output a signal to solve a certain problem based on the input signals it received. The interconnected simple processing units have adjustable gains that are slowly adjusted through iterations influenced by the input output patterns given to the ANN (Fausett, 1994). Most of the neural networks in the brain (especially cortex) are formed by two-dimensional layers of cellular modules that are densely interconnected between them. The response signals of these areas are obtained in the same topographical order on the cortex in which they were received at the sensory organs (Aldasoro & Aldeco, 2000).