IMAGE PROCESSING OF DIGITAL MAMMOGRAMS FOR BREAST CANCER DETECTION AND CLASSIFICATION

by

AHMAD NABIL BIN MOHD NIZOM

Thesis submitted in fulfilment of the requirements for the Bachelor Degree of Engineering (Honours) (Aerospace Engineering)

June 2018

ENDORSEMENT

I, (student's name) hereby declare that all corrections and comments made by the supervisor and examiner have been taken consideration and rectified accordingly.

(Signature of Student)

Date :

(Signature of Supervisor)

Name : Ir. Dr. Parvathy Rajendran

Date :

(Signature of Examiner)

Name : Dr. Elmi Abu Bakar

Date :

DECLARATION

This thesis is the result of my own investigation, except where otherwise stated and has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any other degree.

_____

(Signature of Student)

Date :

ACKNOWLEDGEMENT

First and foremost, I would give my highest gratitude to Allah The Almighty for allowing me to finish this paper.

I would like to thank my supervisor Ir. Dr. Parvathy Rajendran for guiding me in completing this thesis through her help in the technical aspects and writing aspects. There are multiple instances where my limited knowledge in the field is insufficient and I still managed to solve the problem with the help of my supervisor.

Besides that, I would like to give my gratitude to the all lecturers of the school, my family and my friends for the moral support that I have gained from them throughout the period of completing this research.

TABLE OF CONTENTS

ABSTRACT

Due to an increase in number of breast cancer screening worldwide, development of accurate CAD is needed for tumour detection in mammograms. This study aims to develop an image processing algorithm that can produce lesser errors than human operators. The algorithms to be developed will consist of pre-processing, enhancement and image segmentation. This study also aims to develop an algorithm that uses conversion of greyscale image into RGB as an approach to image processing for greyscale image. For the image processing, the pre-processing is done by removal of artefacts and pectorals muscle using image segmentation and selection by region area and region ID respectively. Then, the process begins with the image enhancement using CLAHE to improve the details and contrast in the image. After that, the greyscale image undergo conversion into RGB by changing the colourmap. The image is segmented based on colour then translated into a circle which centroid is same with the cluster and the number of pixel is same to the tumour detected for comparison with the ground truth data. The accuracy of the algorithm developed in detecting tumour is 94.38% showing that it is relevant for use by radiologists. The algorithm may be developed for application in other field that uses greyscale image as well.

ABSTRAK

Dengan peningkatan jumlah pemeriksaan kanser payudara di seluruh dunia, pembangunan system CAD yang tepat adalah diperlukan bagi mengesan tumor melalui mammogram. Kajian ini bertujuan untuk menghasilkan sebuah algoritma pemprosesan gambar yang mampu mengurangkan kadar kesilapan berbanding pengendali. Algoritma yang akan dihasilkan akan merangkumi langkah pra-pemprosesan, penambahbaikan dan segmentasi imej. Kajian ini juga bertujuan untuk menghasilkan algoritma yang menggunakan penukaran imej skala kelabu kepada imej berwarna RGB sebagai sejenis pendekatan bagi pemprosesan imej skala kelabu. Bagi pemprosesan imej ini, fasa pra-pemprosesan dijalankan dengan membuang artefak dan otot dada menggunakan segmentasi imej dan pemilihan melalui saiz keluasan kawasan dan ID kawasan masing-masing. Kemudian, fasa pemprosesan dimulakan dengan penambahbaikan imej menggunakan CLAHE bagi menambahbaik butiran dan kontras dalam imej yang digunakan. Kemudian, imej skala kelabu yang digunakan ditukar kepada gambar berwarna RGB. Imej tersebut kemudiannya disegmentasi berdasarkan warna dan diterjemahkan menjadi bulatan yang mempunyai centroid yang sama dengan kluster asal dan bilangan piksel yang sama dengan tumor yang dikesan untuk dibandingkan dengan data sebenar. Ketepatan algoritma yang dihasilkan dalam mengesan tumor adalah 94.38% tepat menunjukkan ia sebagai relevan untuk digunakan oleh pakar radiologi. Algoritma ini boleh dibangunkan dan digunapakai dalam masalah lain yang menggunakan gambar skala kelabu juga.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3-D	: Three dimensional
AHE	: Adaptive histogram equalisation
CAD	: Computer-aided diagnosis
CDF	: Cumulative distribution function
CIS	: Carcinoma in situ
CLAHE	: Contrast-limited adaptive histogram equalisation
ICD	: International Statistical Classification of Diseases and Related Health Problems
FDA	: U.S.A Food and Drug Administration
HSV	: Hue-Saturation Value
kVp	: Peak kilovoltage
RGB	: Red-Green-Blue colourspace
ROI	: Region of interest
RSD	: Reflex sympathetic dystrophy
WHO	: World Health Organisation

LIST OF SYMBOLS

β_0 : Binary index for original image

$\beta_{x,y}$: Binary index for pixel in at coordinate (x,y)

β_τ : Binary index for segmented tumour mass

β_λ : Binary index for borderline of segmented tumour mass

I : Pixel intensity value

I_0 : Original pixel intensity value

$I_{\max}$: Maximum pixel intensity value

$I_{\min}$: Minimum pixel intensity value

I_{new} : New pixel intensity value

$I_{x,y}$: Pixel intensity value at coordinate (x,y)

x_{new} : New x coordinate of pixel

n_x : Number of pixels on x-axis

n_I : Number of classes of pixel intensity in the selected histogram range

χ : Current x-coordinate of pixel

x_R : Current x-coordinates of pixels in selected region R

y_R : Current y-coordinates of pixels in selected region R

τ_P : True positive pixels

τ_N : True negative pixels

ξ_P : False positive pixels

ξ_N : False negative pixels

r : Radius

A_{tot} : Total pixel count in whole image

A : Area/pixel count in a region

A_I : Area/pixel count in the selected range of pixel intensity

$A_{\max}$: Largest area/pixel count among regions in the image

W : Weightage for selected pixel intensity range

W_b : Weightage of background region

W_f : Weightage of foreground region

μ : Mean pixel intensity for selected pixel intensity range

σ : Pixel intensity variance for selected pixel intensity range

σ_b : Pixel intensity variance of background region

σ_f : Pixel intensity variance of foreground region

σ_{wc}^2 : Within class pixel intensity variance for selected pixel threshold value

T_0 : Initial threshold

T : Global threshold value

R : Region ID

p_I : Probability of occurrence for pixel with intensity I

$cdf_{I,0}$: Cumulative distribution function for pixel with intensity I in initial image

χ : Contrast limiter for CLAHE

ρ_{global} : Maximum index in RGB matrix

ρ_{red} : Maximum index in R matrix

ρ_{blue} : Maximum index in B matrix

ρ_{green} : Maximum index in G matrix

CHAPTER 1

INTRODUCTION

This chapter introduces the project and presents the current existing problems regarding breast tumor detection and the objectives of the research.

1.1 Problem Statement

The number of people going to frequent screening for breast cancer is increasing every year. However, doing frequent screening also increases the risk of having a breast cancer.(Grabler, 2017) As a solution, the need of frequent screening can be reduced by using CAD. Besides that, the dependency on radiologists can be reduced and remove operator's error.

Until now, there are many research being conducted in order to develop a working CAD system on various machines for diagnosing breast cancer. The accuracy of the system must be comparable to manual radiologist reading using film mammography.

A study by Van Djick found that 13% of all cases were found to be errors.(Van Djick, 1993) To ensure that the CAD system created is relevant to be used in the field, the errors produced must be less than said percentage.

One of the most important part of the whole image processing of breast cancer is the image segmentation where the ROI is extracted from the image after enhancement in the previous steps. For image segmentation, the common approach is to find a value that acts as a threshold that will segment the image into background and foreground

components. However, due to limited dimensions of the greyscale colourmap, segmented pixels may be put in the wrong group.

1.2 Objectives

The research presented in this paper is conducted with regards to the objectives which are developing pre-processing algorithm that removes noise, developing image enhancement algorithm that improves clarity of image details and developing color-based segmentation with accuracy higher than 87%. These objectives are important to drive the research in a way that a specific end result can be obtained to determine whether the research meets its purpose or not.

The first step in image processing is the pre-processing stage. In this stage, the pectorals muscle of the digital mammogram is removed. The aim of the first part of the algorithm developed in this research is to properly segment the pectorals muscle and remove it from the mammogram to reduce the false positive error. Besides the pectorals muscle, certain images from the dataset contains artefacts which can also be removed to reduce the false positive error.

The next step is image enhancement. Image enhancement is the phase where an image is enhanced to improve the accuracy of image segmentation phase. This study aims to develop an algorithm where the enhancement stage includes contrast enhancement and converting the greyscale image into an RGB image in which these two operations can help increase the accuracy of the image segmentation.

The contrast enhancement is to be optimized to reduce overamplification of the pixel intensity index while still increasing the contrast from its initial value.

Each pixel in a greyscale image can be represented in a single dimension while each pixel in an RGB image can be represented in 3-D index. By converting the image into RGB, this allows the image segmentation to operate in a more specific manner. Therefore, this research aims to develop an image enhancement method that allows the application of conversion from greyscale image into RGB for image segmentation.

Image segmentation separates pixels into foreground and background objects. The segmented image is presented in binary form where the white region is in the foreground while the black region is in the background.

In this study, the ground truth of the dataset obtained from mini-MIAS database used for image processing is compared with the result to obtain the accuracy of the image segmentation. The aim of the segmentation process is to ensure the accuracy of the segmentation is above 87% to be relevant for usage by radiologists in diagnosis of mammograms.

CHAPTER 2

LITERATURE REVIEW

This chapter presents the information found from previous research done regarding the subject of research and discussion in this paper that can aid in understanding the research better.

2.1 Image Processing

Image processing is a general term covering all forms of processing of image data.(Fisher, 2004) It is a process or a series of process that involves performing one or more operations that can alter, enhance or extract information from an image. There are two types of image processing which are digital image processing and analog image processing. This paper discusses about digital image processing on the dataset of digital mammograms.

Various techniques of digital image process was developed early since the 1960s by multiple well-known research institutes including Massachusetts Institute of Technology, University of Maryland and Bell Laboratories. The application of the research by this institute includes satellite imagery, wire-photo standards conversion, videophone, photograph enhancement, character recognition and medical imaging.(Rosenfeld, 1969)

The potential of digital image processing to utilise complex algorithms allows the production of more sophisticated performance and results when doing simple tasks where the methods used are impossible to achieve with analog image processing. These

processes includes image classification, extraction of features, image projection, pattern recognition and multi-scale signal analysis.

2.2 Image Segmentation

Image segmentation is a process where digital images are partitioned into segments of pixel in order to simplify and alter an image representation into a version that is significantly easier to analyse.(Barghout, 2004)

Typically, the purpose of image segmentation is to locate a desired object in an image and its boundaries. The segmented results collectively covers the whole image but each partitioned segment can be extracted to be viewed more clearly. In each segment, the pixel consist of similar characteristics such as colour, texture or intensity. Segments that are adjacent often are significantly different from each other when considering the same characteristic.(Stockman, 2001)

In this study, the segmented image is represented as binary image. A binary image is digital and has only two possible values for every pixel where the most commonly used colour is black and white. These two colours contrast with each other and are categorised as foreground and background colours.(codersource, 2005)

An application of image segmentation is in medical imaging for diagnosis purpose and locating tumours.(Wu, 2014) There are multiple studies that uses image segmentation in various approaches to extract tumour region from a digital image.

Methods used in image segmentation includes thresholding, histogram-based methods and various other methods. Each method has its own specialty where different types of dataset can be segmented better with different type of image segmentation

process. Thresholding is by far the simplest method of image segmentation that is used to convert greyscale image into binary images. The methods of obtaining the threshold varies, including Otsu's method, balanced histogram thresholding, k-means clustering and recently, multi-dimensional fuzzy rule-based non-linear threshold.(Batenburg, 2009)

2.3 Medical Imaging

For some clinical analysis and medical intervention, medical imaging has been a selected technique where the visual representations of a patient's interior and function of organs, also known as physiology, is created. There are often datasets establish in medical imaging databases to provide a clear ground truth on how to differentiate normal anatomy and physiology to its abnormal counterparts. There has been approximately 5 billion medical imaging studies conducted worldwide up to 2010 alone and the number is increasing day by day.(Roobottom, 2010) Medical imaging has been very abundant that in 2009, the United States of America's National Council on Radiation Protection and Measurements (NCRP) released a report that for the year 2006, 50% of ionizing radiation exposure in the United States of America is from medical imaging alone.(Schauer, 2009)

Most medical imaging procedures are non-invasive which means that a patient does not require to have any incisions or having medical instruments introduced directly on their body.

The two categories of medical imaging are radiology and visible light medical imaging. Results from radiology medical imaging are presented as images by using a special equipment to interpreted by a radiologist. Examples of these medical images are x-ray and mammograms. On the other hand, visible light medical imaging involves

taking still pictures or recording a video that can be seen without needing special equipment. Visible light medical images are obtained in processes like colonoscopy. The application of medical imaging are mainly in medical science fields such as neuroscience, cardiology, psychology, psychiatry, biomedical engineering and medical physics. However, recent development of medical imaging made it relevant for scientific and industrial application including fields like computer science. (James, 2014)

An example of radiology medical image and visible light medical image is shown in Figure 2.3.1 and Figure 2.3.2 respectively.

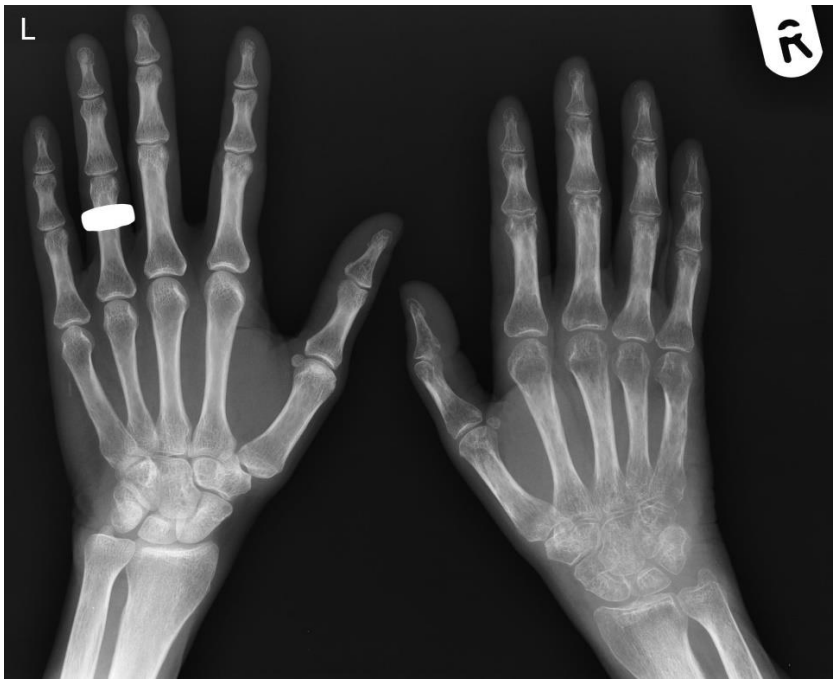

Figure 2.3.1 : Image obtained from x-ray showing reflex sympathetic dystrophy (RSD) (image from www.radpod.org labelled for non-commercial use)

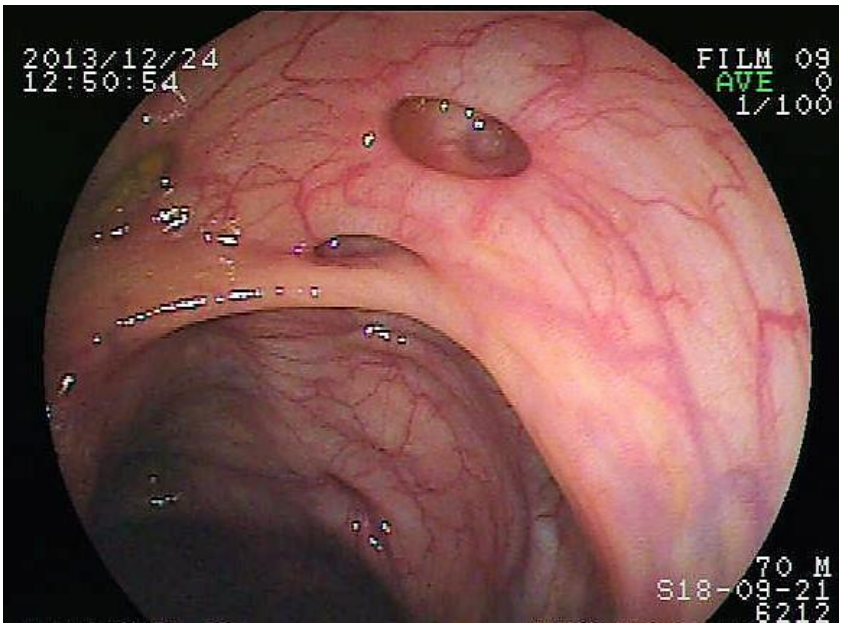

Figure 2.3.2 : Image obtained from colonoscopy showing diverticulosis condition (image from wikipedia.org/wiki/Diverticulosis labelled for non-commercial reuse)

2.4 Digital Mammograms

Mammography is a process where low-energy X-rays is used to examine the human breast for diagnosis and screening producing a medical image known as mammograms. The peak voltage transmitted across the X-ray machine cathodes are presented in peak kilovoltage (kVp) where the recommended value is 19 kVp, increasing up to 29 kVp as the thickness of breast increase from 4 cm to 8 cm.(Fahrig, 1992)

Yearly screening mammography is recommended by the American College of Radiology and American Cancer Society starting at age 40.(American Cancer Society, 2014) Each screening session must be done with long period of intervals. A report by the US Preventive Services Task Forces in 2016 suggests that frequent mammography screening may cause an increased risk of breast cancer due to radiation experience by the patient.(UPST Force, 2016)

Traditional mammography methods uses film to present the result, similar to common x-rays. Current technology uses digital mammography and presents the result as greyscale image in digital form rather than on film. The accuracy of digital mammography is similar to film mammography where digital mammography has higher positive biopsy rate but the difference is not significant.(Lewin, 2001) The insignificant difference in accuracy between the two methods show that digital mammography is also dependable in screening for breast cancer detection.

Digital mammograms allow the use of computer algorithms to process the produced image to extract information from the mammogram. The first digital mammography system to receive an approval by the U.S. Food and Drug Administration (FDA) back in the year 2000. (Maitra, 2011) The images obtained can be enhanced and manipulated to help with identifying the results. The convenience of being able to store and retrieve digital mammograms electronically also is considered one of the reason that makes digital mammography more favourable to some clinical practitioners.

2.5 Computer Aided Diagnosis (CAD)

CAD is a system that uses computer algorithms to aid operators in interpreting medical images. It is basically a pattern recognition-based algorithm that scans a medical image for suspicious anomalies in the structures.

The CAD data are often analysed through processes including pre-processing, segmentation, analysation and extraction of ROI and evaluation. Currently, the application of CAD includes in the diagnosis of breast, lung, colon and prostate cancer, pathological brain detection, Alzheimer's disease, bone metastases, coronary artery disease, congenital heart defect and diabetic retinopathy.

As an alternative method in automatic classification of breast cancer, CAD is believed to be able to reduce the errors due to ultrasound imaging being operator dependent.(Singh, 2016) In 2008, a study by Gilbert, et. al., has proven that CAD may improve the screening process and proposes CAD as a substitute to independent double reading as both methods produce similar accuracy.(Gilbert,2008)

Some other types of cancer that may benefit from CAD are lung cancer and colon cancer. According to the American Cancer Society, early detection in cancer cases are very valuable for the patients as it may reduce death risks, providing survival rate at 47% for lung cancer cases.(American Cancer Society, 2006) For colon cancer, CAD is used for detection of colorectal polyps which are small growths in the inner lining of colon. A study by Suzuki proves that CAD has a significantly high accuracy in detecting colon cancer and can be used as a tool for colon cancer diagnosis from images obtained from CT colonoscopy.(Kenji, 2008)

2.6 Neoplasms/Tumour

Body cells can experience abnormal growth which often forms a mass but not always. Neoplasm is a mass formed by abnormal growth of tissue which is also commonly known as tumour. (Birbrair, 2014)

The 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD) which is a medical classification list commonly known as ICD-10 produced by the World Health Organization (WHO) classifies neoplasms into four main categories which are benign neoplasms, in situ neoplasms, malignant neoplasms also known as cancer and neoplasms of unknown behaviour.(WHO, 2004)

A tumour is either benign, potentially malignant or malignant depending on the diagnosis. Cases of benign tumours are such as osteophytes, uterine fibroids and melanocytic nevi which is known as skin mole. These tumours are localised and circumscribed and do not transform into cancer.(Ingraffea, 2013)

Potentially malignant neoplasm are cases such as carcinoma in situ or most often known as CIS. (Lamm, 1992) These neoplasms may develop become malignant but in this stage, they are harmless and non-invasive.

Cancer cells are malignant neoplasms which invades and destroys surrounding tissue. As a malignant neoplasm grow, it may form metastasis in which it will spread to other regions of the body. Failure to provide responsive treatment may be fatal to a patient. (Pani, 2010)

There are also cases of secondary neoplasm which refers to malignant tumour that occurs from metastasis of the primary cancer tumour or tumour of increased frequency following cancer treatments such as radiotherapy or chemotherapy.(Powe, 2010) Besides that, a rare case of cancer with unknown origin can occur where a cancer tumour come from metastasis but the primary site of the cancer is unknown.

Neoplasms are known to be heterogenous and contain more than one type of cell. However, the initiation and continued growth of these tumour cells are often dependent on a singular population of neoplasms. Being derived from the same cell, these cells are presumed to be clonal.(Nowell, 1976) A proof of these cells being clonal is that they carry same genetic anomaly.

Clonality is proven by the amplification of a single rearrangement of their immunoglobulin gene (for B cell lesions) or T cell receptor gene (for T cell lesions) for cases of neoplasms in the lymph system or known as lymphoid neoplasms such as lymphoma and leukaemia. As a means to properly identify a lymphoid cell proliferation as neoplastic, the demonstration of clonality is now considered to be necessary.(Lee, 1995)

2.7 Breast Calcification

In many cases of breast pathologies, calcium is often deposited at sites of cell death or in association secretions or hyalinized stroma. The deposition of calcium will result in pathologic calcification. For example, small-sized, irregular, linear calcifications may be seen, via mammography, in a ductal carcinoma-in-situ to produce visible radio-opacities. (Kumar,2014)

Two types of breast calcifications have been described from a chemical point of view. (Oyama, 2002) Type I calcifications appear as colourless, birefringent crystals made of calcium oxalate dihydrate (CaC_2O_4-$2H_2O$).(Martin, 1999) This type of calcification is considered benign. Type II calcifications are carbonated calcium phosphate apatite that may exists as either benign or malignant. A study by Baker in 2010 show a significant relationship between the level of carbonate of this apatite and cancer.(Baker, 2010)

2.8 Breast Cancer

Breast cancer develops from breast tissue into malignant neoplasms. Symptomps of having a breast cancer may include a lump in the breast, a change in shape of breast,

dimpling of the skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin. (Gamagami, 1996)

Accounting for greater than 1.6% of the total deaths, breast cancer is one of the major causes of cancer deaths in women worldwide.(Singh, 2016) Multiple risk factors for developing breast cancer include being female, obesity, alcohol consumption, lack of physical exercise, early age during first menstruation, child-bearing at later age or lack thereof, ionizing radiation, hormone replacement therapy during menopause, old age, prior history of having breast cancer and genetics.(PATE Board, 2014)

For early detection of breast cancer cell, one can go for screening or clinical exam. A clinical or self breast exam involves feeling the breast to find abnormalities. Clinical breast exams are performed by health care providers and practicioners, while self-breast exams are performed by the person onto themselves. (CfDCa Prevention, 2017) However, studies do not support both types of exams as it is highly likely that when the mass has grown big enough to be able to be felt, it must have been growing for several years making it can be found without an exam.(UPST Force, 2016)

The recommended approach in diagnosing breast cancer is through mammography screening. There are two methods of doing mammograms, the first one being general mammograms where the entire breast is focused on while diagnostic mammogram only focus on a specific region or lump detected. (American Medical Association, 2007)

Breast cancer may be treated with surgery, followed by chemotherapy or radiation therapy or both. Multidisciplinary approach is more preferable than single

disciplinary approach.(Saini,2011) The medication available for breast cancer patients include hormone blocking therapy, chemotherapy, and monoclonal antibodies.

CHAPTER 3

METHODOLOGY

This chapter presents the programming flow chart and mathematical models applied and used to develop the image processing algorithm from the pre-processing step to the segmentation.

3.1 Programming Flow Chart

Figure 3.1.1 shows the flow chart of the whole image processing algorithm including the pre-processing steps taken prior the image processing itself.

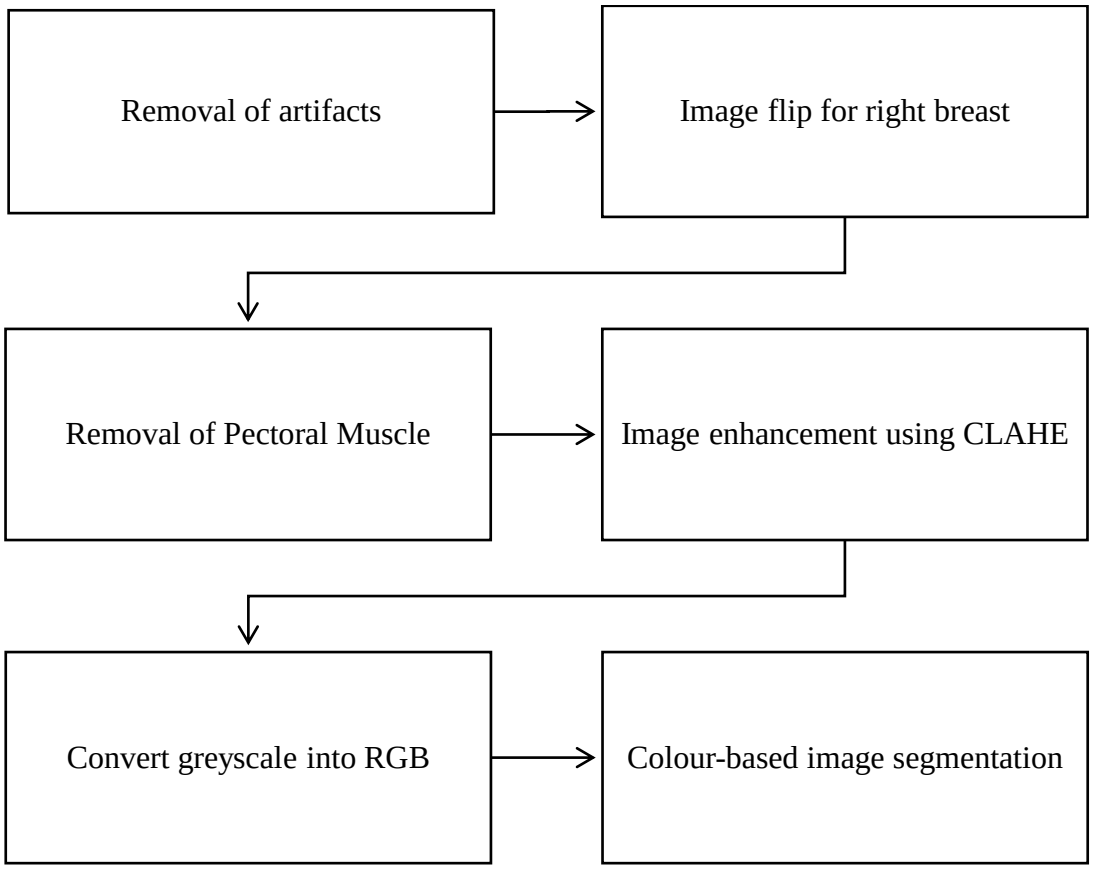

Figure 3.1.1 : Flow chart of step-by-step image processing algorithm

The algorithm starts with removal of artefacts and pectoral muscle as the pre-processing steps to reduce the probability of false positives from occurring. Then the image is enhanced and converted to RGB in HSV colourmap to increase accuracy of image segmentation process.

3.2 Pre-processing Algorithm

The process is started with a series of three pre-processing algorithms. The first step is a compulsory which is artefacts removal, followed by an optional step which is image flipping for right breasts, and then followed by another compulsory step which is pectorals muscle removal.

3.2.1 Artefacts Removal

The first compulsory pre-processing algorithm to reduce error is the removal of artefacts and labels from the digital mammogram image. Figure 3.2.1 shows a sample of digital mammogram with an artefact in the image. The selection of area is done using equation (3.2.1) that translate the greyscale image into binary image. Besides removing the artefact, this method also remove noise and errors in the image.

$$\beta_0 = \begin{cases} 1, & I_0 > 0.15 \times I_{\max} \\ 0, & I_0 < 0.15 \times I_{\max} \end{cases} \tag{3.2.1}$$

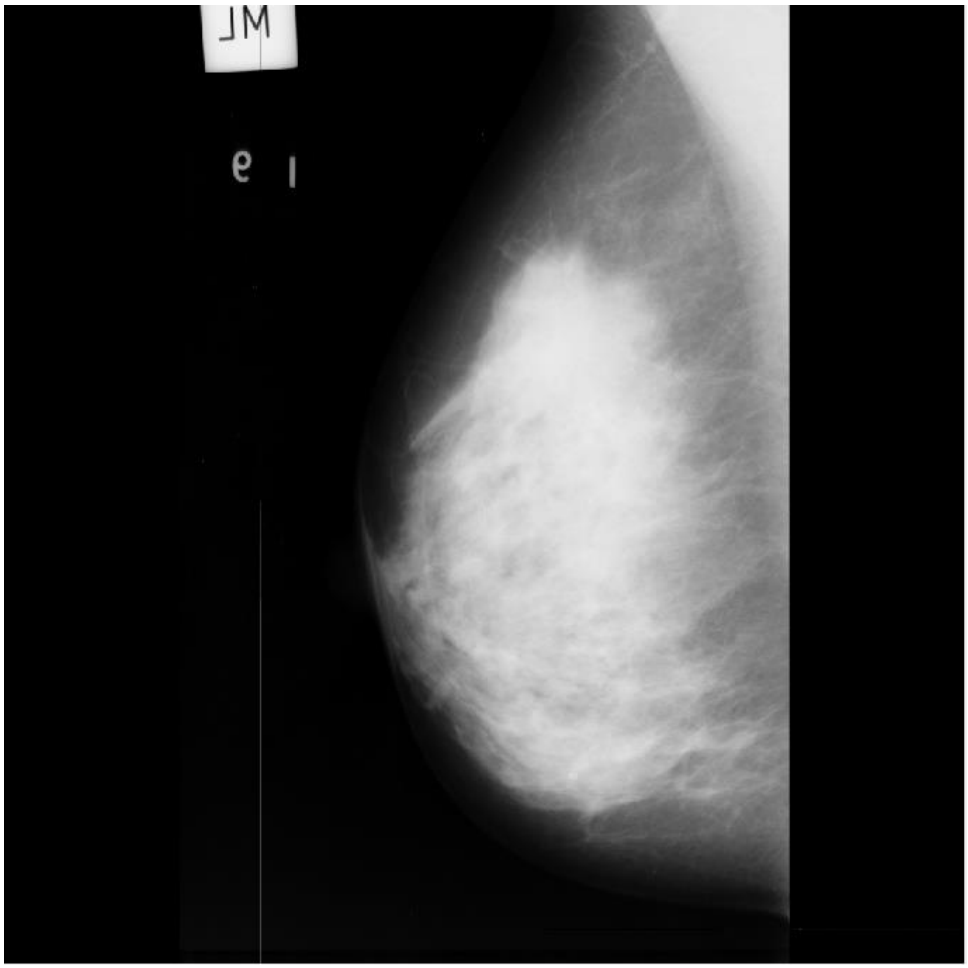

Figure 3.2.1 : Image mdb003 from Mini-MIAS database

The binary image produced contains multiple regions of masses of different sizes. The breast region and pectoral muscle region are combined due to having higher intensity than the threshold. The largest region of the binary image is then selected and retranslated as a greyscale image as a mask. The selection method follows equation (3.2.2) as shown below.

$$I_{new} = \begin{cases} \beta_0 \times I_0, & A = A_{\max} \\ 0, & A < A_{\max} \end{cases} \qquad (3.2.2)$$

3.2.2 Image Flipping

After removing the artefacts, the image is checked whether the pectorals muscle region is at the left or the right side of the image. Left breast images have the pectorals muscle on the left side while the right breast images have the pectorals muscle on the left side of the image. These pectorals muscle region always remain on either side and never exceed the middle of the image. Therefore, the digital mammogram image will only be flipped across the vertical axis if there is no pixel value up until the said point. This operation is needed to simplify the pectoral muscle removal model and follows the equation (3.2.3). By flipping the image only across the vertical axis, the orientation of each image to be processed become similar.

$$x_{new} = \begin{cases} (n_x + 1) - x, \sum_{x=1}^{0.4n_x} I_{x,1} = 0 \\ x, \sum_{x=1}^{0.4n} I_{x,1} > 0 \end{cases} \quad (3.2.3)$$

3.2.3 Pectoral Muscles Removal

The pectoral muscle removal is done via segmentation as well, applying the Otsu's Method for thresholding. As we know, binarization separates pixels that is higher than or equals to the set threshold as the foreground. Otsu's method is an algorithm that iterates the variance in the foreground and background from the number of pixel of the histogram components in each class. This process is shown in equation (3.2.4) to (3.2.6) which applies to both foreground and background region but must be calculated separately. Weightage is a measure of how significant is the influence of the foreground

or background class in determining the variance within class. The mean pixel intensity value dictates the average pixel intensity in either foreground or background class. The variance shows how varying each pixel's intensity value compared to the mean intensity in either foreground or background. Basically, a highly dynamic range of pixel intensity increases the variance.

$$W = \frac{\sum_{i=1}^{n_I} A_i}{A_{tot}} \tag{3.2.4}$$

$$\mu = \frac{\sum_{i=1}^{n_I} \left(I_i \times A_i \right)}{A_I} \tag{3.2.5}$$

$$\sigma = \frac{\sum_{i=1}^{n_I} \left(\left(I_i - \mu \right)^2 \times A_i \right)}{A_I} \tag{3.2.6}$$

From the variance and weightage of both foreground and background class, the variance within class can be found by equation (3.2.7). Then, the threshold for the iteration with smallest within class variance is selected as the global threshold value.

$$\sigma_{wc}^2 = W_b \cdot \sigma_b^2 + W_f \cdot \sigma_f^2 \tag{3.2.7}$$

In this study, the background is a major part of the image itself, causing the Otsu Method to produce a very low global threshold. Therefore, the threshold obtained is multiplied by half of the difference between the threshold and the maximum pixel intensity to increase the threshold. The process can be shown in equation (3.2.8) as follows.

$$T = T_0 + \frac{\left(T_{\max} - T_0\right)}{2} \qquad (3.2.8)$$

Using the threshold value, T, the image is binarized creating two disconnected region which are the breast region and pectoral muscle region. Since the image is oriented so that the pectorals muscle is always on the left side of the image, the first mass is removed using the following equation (3.2.9). The region ID, R , is assigned starting by integers starting from 1 where the first region is the region with pixel at smallest x coordinate and largest y coordinate. Essentially, this means that the region R_1 is the first from the top left corner.

$$\beta_{x_R, y_R} = \begin{cases} 0, R = R_1 \\ \beta, R > R_1 \end{cases} \qquad (3.2.9)$$

After removal of the pectoral muscle region from the binary image, the image is converted back into the greyscale image which now has its pectorals muscle removed using equation (3.2.10) which is essentially similar to the first equation in the piecewise equation (3.2.2). The image obtained can be further processed in the next steps. The result of this masking is a completely black background with ROI in the same intensity as the original image.

$$I_{x,y} = \beta_{x,y} \times I \qquad (3.2.10)$$

3.3 CLAHE Image Enhancement

CLAHE is a method that equalizes histogram locally allowing the each local region to have limited contrast among each of it component pixels while increasing its distinction to the neighbour regions making the edges of each region more clear. This is very helpful in enhancing the gradient edges of some regions which may cause some part of the mass that should be segmented to be missed.

Being an improved form of AHE which is obtained from histogram equalisation algorithms, to understand the process of CLAHE, the process of histogram equalisation and AHE must first be understood.

Histogram equalisation is a contrast adjustment process commonly used in image processing that depends on the histogram of the image. In this study, the digital mammogram is a discrete greyscale image. The histogram level of each pixel intensity value can be obtained from the equation (3.3.1) where the p_I does not only present the probability of occurrence but also the histogram level.

$$p_I = \frac{A_I}{A_{tot}}, 0 \le I < I_{\max} \qquad (3.3.1)$$

From the p_I , we can determine the cumulative distribution function by using the equation (3.3.2). The slope of the cumulative distribution function shows the slope of the histogram itself. If the cumulative distribution function is linear and has a constant slope, it means that the histogram is flat while a histogram with normal distribution will

correspond with a cumulative distribution function slope that exponentially increase approaching the peak of the histogram and decay after the peak of the histogram.

$$cdf_{I,0} = \sum_{j=0}^{I} p_j \qquad (3.3.2)$$

To transform the image into a histogram equalised version which has a flat histogram, the cumulative distribution function must be linear. Therefore, the transformed cumulative distribution function can be denoted as equation (3.3.3), where K is a constant.

$$cdf_{I,new} = I \cdot K \qquad (3.3.3)$$

In this case, the cumulative distribution function is always increasing because there is no negative values in the histogram. Therefore, the inverse cumulative distribution function can be defined by equation (3.3.4)

$$cdf_{I,new}^{-1} = cdf_I(K),\ 0 \le K \le 1 \qquad (3.3.4)$$

From this, the new histogram can be mapped back into its range by equation (3.3.5).

$$I_{new}' = I_{new} \cdot \left(\max\{I_{old}\} - \min\{I_{old}\}\right) + \min\{I_{old}\} \qquad (3.3.5)$$

Adaptive histogram equalisation or AHE uses histogram equalisation methods in localised manner. The image is divided into M-by-N grids with its centre known as grid

points. Each pixel histogram is measured in comparison to its nearest adjacent grid points. This approach allows histogram equalisation to be done without sacrificing the contrast of images with the background and ROI being the most abundant in the extremum as seen in most digital mammograms. However, in regions of noise, the AHE will overamplify the noise in the region. For example, if a single pixel has maximum intensity value, while the neighbouring pixel in a local 6-by-6 grid pixel intensity of 0 due to it being a background region, then the whole region will average out to become a grey area.

By using CLAHE, a contrast limiter is set so that overamplification cannot happen. This process can be modelled by using the following equation (3.3.6) and (3.3.7). The contrast limiter is the maximum difference between the brightest and darkest pixel.

$$I_{new} = \begin{cases} I_{new}, |I_{new} - I_{old}| \leq \chi \\ I_{CLAHE}, |I_{new} - I_{old}| > \chi \end{cases} \tag{3.3.6}$$

$$I_{CLAHE} = \begin{cases} I_{old} - \chi, I_{new} - I_{old} < 0 \\ \chi - I_{old}, I_{new} - I_{old} > 0 \end{cases} \tag{3.3.7}$$

An example result of image that has been enhanced by CLAHE is shown in Figure 3.3.1 and Figure 3.3.2. Figure 3.3.3 shows the results of using traditional contrast enhancement for comparison. As can be seen in Figure 3.3.3, the enhancement shows lesser details in the edges compared to Figure 3.3.2. Besides that, CLAHE does not drastically modify the histogram to approach its extremum as much as traditional contrast enhancement which creates higher pixel count in the extremum.

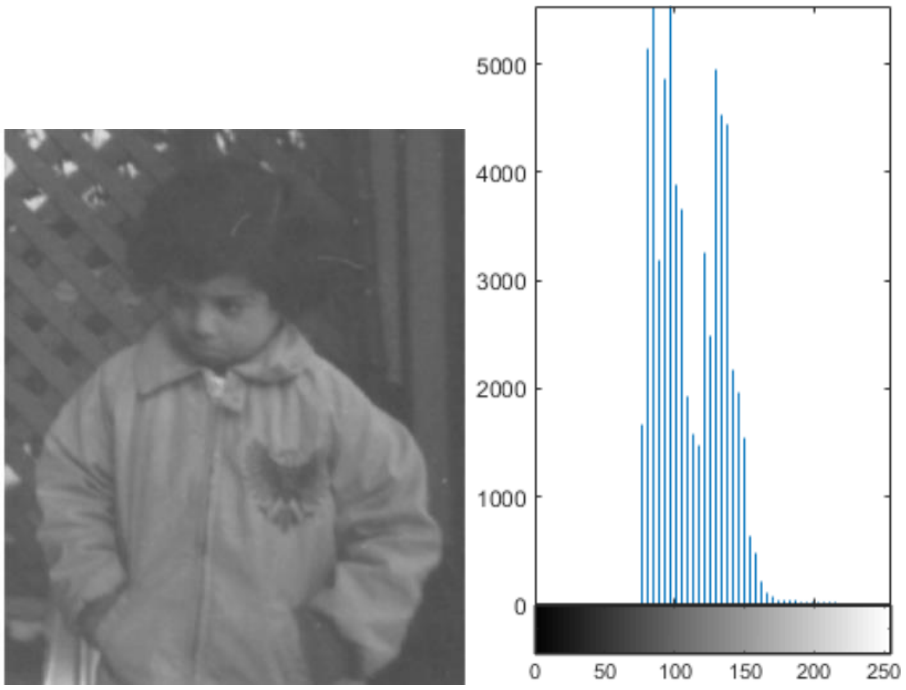

Figure 3.3.1 : Sample greyscale image in its original form (*mathworks.com*, accessed 1/5/2018)

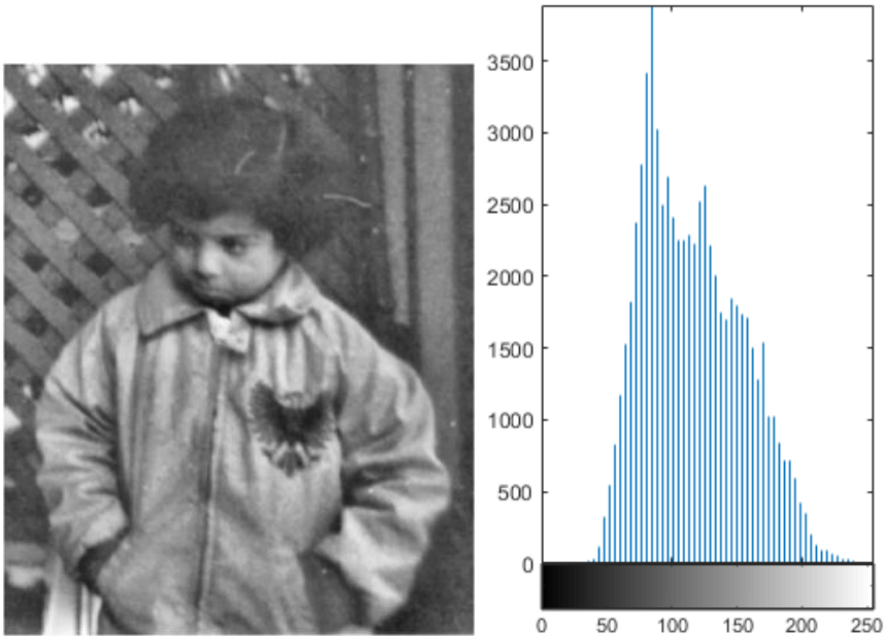

Figure 3.3.2 : Sample image enhanced using CLAHE (*mathworks.com*, accessed 1/5/2018)