COMPARATIVE STUDY OF IMAGE OUTLIER DETECTION USING HEXAGONAL LATTICE AND RECTANGULAR LATTICE BASED ON SELF-ORGANIZING MAP

WAN SAIFUL 'AZZAM BIN WAN ISMAIL

UNIVERSITI TEKNOLOGI MALAYSIA

COMPARATIVE STUDY OF IMAGE OUTLIER DETECTION USING HEXAGONAL LATTICE AND RECTANGULAR LATTICE BASED ON SELF-ORGANIZING MAP

WAN SAIFUL 'AZZAM BIN WAN ISMAIL

A project report submitted in partial fulfillment of the
requirement for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

OKTOBER 2009

" *To my beloved family and friends, thanks for being there throughout this journey* "

ACKNOWLEDGEMENTS

Alhamdulillah, praise to the Almighty, Allah S.W.T for giving me a strength and passion that I was able to complete my master project within the given time. On the other hand, not forgotten to the individual that becomes a backbone for the completion of this project. First and foremost I would like to take an opportunity to thank my supervisor *Dr Siti Zaiton Binti Mohd Hashim* for her guidance and support to make sure that this project done smoothly. Also a very thank you to the examiners *PM Dr Puteh Binti Saad* and *Dr Abdullah Bin Bade* whose willing to give a comment and advise in order to improve the quality of this project.

My special thank you also to my beloved family especially my parent for their support and encouragement when I need the most. Beside, also not forgot to my entire friend for their cooperation for helping me in order to complete my project. Last but not least, I could express more thanks to late *PM Md Noor Bin Mat Sap* for his brilliant ideas to make sure this project come true, *Alfatihah*.

ABSTRACT

From the past of the year, Mammogram has drawn an attention for the clinical analysis where it play an important roles for the extraction of information that can be use in diagnosing and treatment disease. According to the significant use of the mammogram for diagnosing, the aim of this project is to detect the outlier from the mammogram images of breast through the clustering method using one of the popular and widely use in Artificial Neural Network Clustering which is Self-Organizing Map (SOM). The outlier indicate as the observation that is far from the rest of the rest of data where it can represent that the data either the unusual data or noise that is important for the further analysis. But, before proceeding to clustering process for the outlier detection, the image must be preprocessing first because the source of the images usually provides with different size and quality where it affect the accuracy of the analysis. Image preprocessing involves the process of crop the region of interest, enhanced the image, remove the noise, and do normalization. Once the image has been preprocessing, the data of the image is extract using Non-negative Matrix Factorization (NMF). NMF has been proven as a powerful method for non-negative data such as image and document. The data that have been extracting from the image using NMF method, the extracted data is apply to the SOM technique to cluster the similarities of the data and at the same time can detect the outlier which is refer to the data that is not in any of clustering. Two type of lattice which is rectangular and hexagonal lattice will use for the training process and compare to find the lattice that can produce the best result.

ABSTRAK

Sejak kebelakangan ini, *Mammogram images* telah memberikan perhatian yang khusus kepada analisis klinikal dimana ia telah memainkan peranan yang penting kepada penguraian maklumat yang boleh digunakan untuk mengdiagnos penyakit. Disebabkan itu, tujuan projek ini dijalankan adalah untuk mengenalpasti data asing daripada sekumpulan *mammogram* dengan menggunakan metod yang sering digunakan didalam *Artificial Neural Network Clustering* iaitu *Self_organizing Map* (SOM). Data asing dapat dilihat melalui kewujudannya yang jauh daripada data-data yang lain dimana ia meberi sesuatu makna samaada maklumat yang belum dikenalpasti lagi mahupun *noise* yang perlu diketahui untuk proses analisa seterusnya. Tetapi, sebelum meneruskan proses *clustering* ini, imej tersebut mestila melalui pra-pemprosesan terlebih dahulu kerana imej-imej yang didapati biasanya mempunyai pelbagai jenis saiz dan kualiti yang berbeza yang akan memberi kesan kepada ketepatan analisa. Langkah-langkah yang terlibat didalam pra-pemprosesan imej termasuklan *crop* bahagian yang dikehendaki, pembaikan imej, pembuangan *noise*, dan juga penormalan. Selepas menjalani pra-pemprosesan, data-data didalam imej tersebut akan diuraikan melalui penggunaan *Non-Negative Matrix Factorization* (NMF). NMF telah terbukti sebagai method yang bagus bagi data bukan negatif seperti imej dan dokumen. Selepas data-data telah diurai daripada imej, data-data tersebut akan digunakan didalam SOM untuk proses *cluster,* data-data yang mempunyai persamaan dan pada waktu yang sama mengenalpasti data asing yang merujuk kepada data-data yang tidak berada dimana-mana *cluster*. Pengunaan dua jenis kekisi yang berbeza iaitu *rectangular* dan *hexagonal* akan digunakan semasa proses pembelajaran dan keputusan akan dibandingkan untuk mencari kekisi yang dapat memberikan keputusan yang lebih bagus.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURES

LIST OF ABBREVIATION

ANN	-	Artificial Neural Network
BMU	-	Best Matching Unit
CT	-	Computed Tomography
NMF	-	Non-Negative Matrix Factorization
PCA	-	Principal Component Analysisi
SOM	-	Self-Organizing Map
qe	-	Quantization Error

LIST OF APPENDICES

CHAPTER 1

PROJECT INTRODUCTION

1.1 Introduction

Within the growth of technologies now days, data imaging have been widely used for the representation in many fields especially in clinical purpose. Data image is a process to create the image of part of human body that can be use for medical procedure such as diagnosis or examine disease or for the medical science for the study of normal anatomy and psychology. Mammogram is one of the data image that has been used widely in the process of diagnosing the disease where usually used for capturing the image of the breast. Mammogram images sometime play an important rule for the diagnosis of disease where it can visualize the internal structure of the breast. It can easily show the abnormalities of the occurrence in the human breast.

Due to the emergence through of using image data, this study is focused on the technique to detect the outlier of the base on the mammogram data image using one of the Artificial Neural Network (ANN) clustering. .The outlier can be shown as a noise of an image or miss interpretation of the data. This process can be done

visually but it became the problem when the number of the outlier must be identify is bigger and also when different observation of reviewer to identify the outlier because different people have different thought. The aim of this project is to cluster the image data using the SOM technique for the outlier detection. In the SOM learning algorithm, two type of the lattice structure for the data mapping will be use and the result of each lattice structure will be compare to get optimization result.

When use an image as a data, this study also will be focus on how to convert an image data to the conventional input data. While converting to the binary form, the image must be preprocessing first such as, image enhancement, cutting the extra background region, remove the noise and normalize to make sure that the image is provided with the only informative data and in the same scalar. The image extracted using Non-Negative Matrix Factorization (NMF) to get the only informative data from the image so that it can be used for the detecting outlier using SOM with the more accuracy and faster computational time for the analysis.

1.2 Background Problem

The growth of digital picture and the need for fully automatic annotation and the increasing of retrieval system have cause the problem of image classification and clustering over the last decade (Krishna Chandramouli, 2007). Image clustering is about grouping the image with the similarities into the same cluster and creates the image classes which provide visualization of image dataset. Clustering has been used in exploratory pattern analysis, grouping, decision making and machine learning-situation including data mining, document retrieval, image segmentation and pattern classification (Sitao Wu and Tommy W.S. Chow 2003). The data image is the powerful tools to visualize and analysis the high dimension hence allows to detect outlier visually.

Recently, the outlier detection is important problem for many domains. Data outlier can have significant impact upon data driven decision. Data outlier could be due to erroneous data or indicate that the data is correct but unusual which is it may indicate that the data is the new abnormal tissues (Brett G. Amidan, Thomas A. Ferryman and Scott K. Cooley , 2004). The identified of potential outlier allowing for the domain expert to investigate the cause. The density based method LOF have been use widely but the complexity of the method is quadratic to size of the dataset, and it may miss the potential outliers when density distributions in the neighborhood are significantly different (Jilin Qu, 2008).

Image also can be used to analysis for the outlier detection due to the widely use of image and helpful for the clinical diagnosing disease and treatment. Mammogram image is one of the methods for diagnosing human body part. Mammogram image is important in order to detect an early stage of breast cancer since from the recent of year; breast cancer is second leading of cancer death today. But using a mammogram it is difficult to interpret because of the small differences densities, difference breast tissue and difference size (J.K. KIM and H.W. Park 1999). Due to the using the image as a data for the analysis, the image usually must be go through the stage of image preprocessing. According to the Rapp, C.S (1997), image preprocessing is the step to modify of image for the improvement of the quality. It is because the image basically provided with different quality and contains the unuseable information. Image preprocessing is important stage especially for the feature extraction and for a mean while it will help to increase the accuracy of diagnosis.

Feature extraction is one of the important stages in the image analysis where it will be used to extract or get the data from the image for the further analysis. It also helps to process of dimensionality reduction because the processing of image requires the large computational algorithm. Non-Negative Matrix Factorization (NMF) has been proven a powerful method for the non-negative data such as image and documentation data W., Liu et al (2007). NMF is a method that is derived from the classical method which is Principal Component Analysis (PCA) where it based

on the representation to find the local features. According to the L., Xu et al (2007), NMF has been prove that produce the effective and accurate feature extraction tool for characterizing the metabolic status of people as belonging to the healthy or diabetes compare to PCA.

SOM is provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane. It is reasonable that two-level approach that is able to cluster by using SOM. The first level is using for the train of the data while the purpose of second level is to clustering the data through SOM. Usually, SOM connect the neurons to each other via two type of lattice which is rectangular or hexagonal topology. The comparison of two topologies is important to the effectiveness for the outlier detection.

1.3 Problem Statement

Based on the previous study, there are many methods for detecting outlier. According to the study, this project is intended to apply the outlier detection through the clustering method using SOM algorithm for the image data. From the study of the SOM algorithm basically two type of lattice which is rectangular and hexagonal lattice are commonly used for the learning process so for this study is to compare the result based on the two lattices in getting the better result. But before can proceed to the image must be go through the preprocessing and feature extraction to get the informative data for the accurate analysis. Non-Negative Matrix Factorization (NMF) has been proven as a one of powerful method for the nonnegative data such as image. So for this project will study the effectiveness of NMF in order to extract the data for the detecting outlier based on clustering using SOM.

1.4 Project Aim

The aim of this project is to compare hexagonal lattice and rectangular lattice based on Self-Organizing Map for image outlier detection.

1.5 Objective

There are several objective of this project that must reach:

1. To perform Non-negative Matrix Factorization for image feature extraction
2. To detect outlier from a series of mammogram images.
3. To compare the using of different lattice (Rectangular and Hexagonal topology) for the mapping in SOM according to the effectiveness of accuracy.

1.6 Scope

The scopes of this project are:

1. Mammogram of breast image is used, taken from https://peipa.essex.ac.uk/info/mias.html because it provides the high quality of mammogram breast image.
2. The outlier detection is in the subset of breast image sample where the outlier is indicating the malignant tumor.

3. SOM will be used for clustering the sample image data to detect the outlier of sample image because from the recent of year SOM is one of the succesfull methods for clustering.
4. Two type of lattice topology which is rectangular and hexagonal were used in the training SOM.

1.7 Significant of Study

This study is focus on the detection of the outlier or the unusual data among the several of brain image data. Basically, the brain consist various type of disease including the normal brain. In order to find the outlier, the image must be organized into the group where the data which is not in any of the group is defined as outlier. Artificial Neural Network for clustering can play the role for grouping the data into a group based on similarity. SOM is the one of the most popular method for the clustering process. The performance of SOM is train using the difference lattice which is rectangular lattice and hexagonal lattice where which lattice will give the best performance.

1.8 Organization of Report

For this project, the report organized into 5 chapters where the first chapter is the introduction of the project then followed by second chapter that discuss the literature review describing about the basic of Non-Negative Matrix Factorization for the feature extraction. Also it covers about the concept of learning in SOM for the

clustering method. In chap 3 it discussed about the methodology that will be used for the overall implementation of the propose solution. In this chapter it will describe the framework of study that must be following to make the process of study done smoothly. In chap 4 the experimental result is presented and analyze and lastly in the chapter 5 the conclusion of the study will be discuss.