UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

PRINCIPAL COMPONENT ANALYSIS – DIMENSIONALITY REDUCTION FOR WRITER VERIFICATION

Rimashadira Binti Ramlee

Master of Science in Information and Communication Technology

2015

PRINCIPAL COMPONENT ANALYSIS - DIMENSIONALITY REDUCTION FOR WRITER VERIFICATION

RIMASHADIRA BINTI RAMLEE

A thesis submitted
in fulfillment of the requirements for the degree of
Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2015

DECLARATION

I declare that this thesis entitled "Principal Component Analysis – Dimensionality Reduction for Writer Verification" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :.....

Name :Rimashadira Binti Ramlee

Date :.....

APPROVAL

I hereby declare that I have read this thesis and my opinion this thesis is sufficient in term of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature :.....

Supervisor Name :Prof Madya Dr. Azah Kamilah Bt Muda

Date :.....

DEDICATION

In the name of Allah, the Most Gracious and the Most Merciful.

I dedicate this work to:

My Parent:

Ramlee Anak Ahad & Rosenani Abdul Jalil.

My Siblings:

Rugieo Ryanzo Ramlee, Rumeilen Perira Ramlee, Ryan Leornard Ramlee, Rashimadini Ramlee & Rashimanshah Ramlee.

ABSTRACT

Writer verification (WV) is a process to verify whether two sample handwritten document are written by the same writer or not. WV also known as one to one comparison process, where the process is more specific which compare one writer to another writer. Therefore, this process needs a unique characteristic of the writer in order to prove the owner of the handwritten document. Basically, different person will have different type of handwriting styles usually it is unique between each other. Furthermore, most of the previous research in handwriting analysis field was used the unique characteristic to represent the individuality of handwriting. A part from that, individuality of handwriting became main issue in this study in order to fulfill requirement of WV process. In previous verification framework of WV the individuality of handwriting was acquired by using feature extraction process. Meanwhile, previous verification framework of WV consists of Preprocessing task, feature extraction task and classification task. In this study, using the previous verification framework are not enough to produce the best result in verification process. This is because the quality of individuality of handwriting that has been acquired is less effective in representing the uniqueness of the writer. Therefore, this study was proposed Dimension reduction technique for acquiring the individual features of the handwritten data henceforth improved the previous verification's framework in order to enhance the verification accuracy. The sample data was taken from IAM online database which this database is the benchmark for handwriting analysis research. Five writers with 3619 instance of images are chosen for the experiment whereas 9 documents of handwriting samples are taken from each writer and more than 50 word randomly divided into training and testing dataset. Both dataset is will be process by Principal Component Analysis which is one of the dimension reduction techniques. PCA was applied after feature extraction process whereas the reduction process will resulted low dimensional of new subspace of data. By using the data resulted by PCA the classification process by random forest was conducted in order to verify the writer of the handwritten document. The individuality representation is implemented by presenting various representations of individual feature into more important feature are selected by using the proposed technique to be used in verifying the writer. Experimental show that the performance of the proposed methods has improved the verification rate of 90.00 % and above overall of the result with the reduction is successful in each data set. However, overall of the result the improved framework still cannot verify 100 % accurately the writer of the handwritten data.

ABSTRAK

Pengesahan Penulis (WV) adalah satu proses untuk mengesahkan sama ada dua sampel dokumen tulisan tangan yang ditulis oleh penulis yang sama atau tidak. WV juga dikenali sebagai 1-1 proses perbandingan, di mana proses ini menjadi lebih khusus yang membandingkan satu penulis kepada penulis yang lain. Oleh itu, proses ini memerlukan ciri unik penulis untuk membuktikan pemilik dokumen tulisan tangan. Pada dasarnya, orang yang berbeza akan mempunyai pelbagai jenis gaya tulisan tangan biasanya ia adalah unik antara satu sama lain. Tambahan pula, sebahagian besar daripada penyelidikan terdahulu dalam bidang analisis tulisan tangan telah menggunakan ciri-ciri yang unik tersebut untuk mewakili keperibadian tulisan tangan penulis. Di samping itu juga, keperibadian tulisan tangan menjadi isu utama dalam kajian ini bagi memenuhi keperluan proses WV. Dalam rangka kerja sebelum ini, keperibadian tulisan tangan telah diperolehi dengan menggunakan proses pengekstrakan ciri. Sementara itu, rangka kerja WV sebelum ini mengandungi pra pemproses tugas, ciri tugas pengekstrakan dan tugas klasifikasi. Dalam kajian ini, dengan menggunakan rangka kerja sebelum ini tidak mencukupi untuk menghasilkan keputusan yang terbaik dalam proses pengesahan. Ini kerana kualiti keperibadian tulisan tangan yang telah diperolehi adalah kurang berkesan dalam mewakili keunikan penulis. Oleh itu, kajian ini telah mencadangkan Teknik pengurangan dimensi untuk memperoleh ciri-ciri individu dari data tulisan tangan serta dapat menambahbaikkan rangka kerja yang digunakan sebelum ini untuk meningkatkan ketepatan pengesahan tulisan tangan. Data sampel telah diambil dari IAM pangkalan data dalam talian dimana pangkalan data ini adalah penanda aras bagi penyelidikan analisis tulisan tangan. Lima penulis dengan 3619 contoh imej dipilih untuk eksperimen manakala 9 dokumen sampel tulisan tangan yang diambil dari setiap penulis dan lebih daripada 50 perkataan secara rawak dibahagikan kepada latihan dan ujian dataset. Kedua-dua dataset tersebut digunakan dalam process PCA yang merupakan salah satu teknik pengurangan dimensi. PCA telah digunakan selepas proses pengekstrakan ciri, dimana process ini akan mengurangkan dimensi data asal kepada data yang baru yang mempunyai dimensi yang lebih rendah. Dengan menggunakan data yang dihasilkan oleh proses PCA, klasifikasi oleh "Random Forest" telah dijalankan untuk mengesahkan penulis dokumen tulisan tangan. Perwakilan keperibadian dilaksanakan dengan mengemukakan pelbagai representasi ciri individu ke dalam ciri-ciri yang lebih penting adalah dipilih dengan menggunakan teknik yang dicadangkan untuk digunakan dalam mengesahkan penulis. Eksperimen menunjukkan bahawa prestasi kaedah yang dicadangkan telah menambahbaik kadar pengesahan peratusan 90.00% dan ke atas dan pengurangan dimensi berjaya dilaksanakan kepada setiap set data. Walau bagaimanapun, secara keseluruhan keputusan rangka kerja yang telah diubah suai masih tidak dapat mengesahkan 100% ketepatan penulis data bagi tulisan tangan.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful. I take this opportunity to express my deep regards and a lot of thanks to my first Supervisor Dr.Azah Kamilah Binti Muda for her guidance, monitoring and constant encouragement throughout the course of this thesis. The blessing, help, and guidance given by her time to time shall carry me a long way in the journey of life on which I am about to start.

I also take this opportunity to thanks my co-supervisor Dr.Burairah Bin Husain, for his support, valuable information and guidance also knowledge that he give me, which help me in completing every task in completing my study.

I appreciation all the staff members of Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka for giving me valuable information in their respective fields. I am grateful for their cooperation during my study.

My special thanks to my parent Ramlee Anak Ahad and Rosnani Abdul Jalil for their constant support and love me all the time. They always stay by my side when I am in disappointment, stressful and bad condition. Their blessings always carry me along the way in the journey of my life.

Finally, I want to express my thanks to my friend that always help me in any condition without paying anything. Especially to NoorAziera Akmal Sukor, Tarisa Makina Kintakaningrum, Lustiana Pratiwi and Satrya Fajri Pratama which always give me a support every time in completing my study .

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WV	-	Writer Verification
DR	-	Dimension Reduction
PCA	-	Principal Component Analysis
LDA	-	Linear Discriminant Analysis
PCs	-	Principal Component
AUMI	-	Aspect United Moment Invariants
WI	-	Writer Identification

LIST OF SYMBOLS

$\propto$ - Threshold level

LIST OF PUBLICATIONS

Ramlee, R., Muda, A. K., and Emran, N. A. 2014. Comparison of Feature Dimension Reduction Approach for Writer Verification. *First International Conference on Advanced Data and Information Engineering*, 285, 92-102.

Ramlee, R., Muda, A. K., and Syed Ahmad, S. S. 2013. PCA and LDA as Dimension Reduction for Individuality of Handwriting in Writer Verification.*13th International Conference on Intelligent Systems Design and Applications (ISDA).*

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.0 Research Background

Nowadays, most of people shift toward typing rather than writing. However handwritten paper document still never been forgotten, because handwriting is a basic of normal human being behavior especially in Forensic Science field. Thus, there always have been situations in which unsigned or anonymous writing on documents were potentially important. Thereby, the provision of proof towards the authorship of such documents has long been an issue particularly in handwriting biometric study (Huber and Headrick, 1999). A part from that, this study is proposed to use Handwriting Analysis approach in order to proof the authorship of handwritten document.

Furthermore, handwriting has long been considered as individualistic and unique especially in examining forgery or authenticity of a document. According to (Zhang and Srihari, 2003a) the individualistic and uniqueness of handwriting rests on the hypothesis that each individuals has unique characteristic of handwriting call as individuality of handwriting. Therefore, this individuality becomes a major issue in this study for implementation of handwriting analysis approach (Srihari, 2002). Apart from that, a lot of application such as criminal justice system will use this individuality to verify the author which involve in crime such as black mail, falsification of the will, etc.

Since the individuality of handwriting is required in handwriting analysis approach, there are several method was proposed in previous research (Bensefia *et al.*, 2004; Zhang *et al.*, 2003b; Zois and Anastassopoulos, 2001). The main purpose of the method is to

obtain the individuality of handwriting from the handwritten data of the writer. Henceforth, this study will propose reduction approach for acquire the individuality of handwriting to overcome the major issue in applying handwriting approach. The basic purpose of reduction concept is facing the irrelevant and redundancy problem that occur in a sample data that can influence the data analysis process. Therefore, this irrelevant and redundancy must be eliminated from the data to enhance the representation of data for analysis purposed (Masaeli *et al.*, 2010).

DR is one techniques of the reduction approach the usually used in data analysis process. DR is chosen because of its facilities which effectively transform the original data into new feature space with reduced dimension (Cunningham, 2007). The new feature space consist only the most significant feature of original data which can be used to represent the individuality of handwriting. Reduction that implemented by DR is based on the dimension of the features vector not by each single feature. This is because DR thought that the entire features are important in representing the sample data. However, this technique will select dimension that consist of the best features vector combination which is most significant feature is in there.

Meanwhile, Writer Verification (WV) method is one of handwriting analysis approach which is the main focus in this study. This study will apply the previous verification framework of WV which consists of preprocessing Task, feature extraction task and classification task (Srihari, 2002). In order to enhance the effectiveness of verification process, this study was planned to inject the proposed DR technique into previous verification framework. Instead of acquiring the individuality of handwriting, DR also capable in enhancement the data representation for classification purposed. Therefore, the classification process run smoothly and data learning process become easier.

1.1 Problem Statements

The main purpose of Writer Verification is to verify the author of handwritten document. In order to implement WV method, this study was facing one major issue which is acquiring the individuality of handwriting in order to improve the verification framework. The necessity of individuality of handwriting is a priority in writer verification process (Bensefia *et al.*, 2004; Srihari, 2002; Zois and Anastassopoulos, 2001). Therefore, obtaining the features that can represent individuality of handwriting became a problem in verification process. The features should be unique among the individual features of the writer and significant to clarify the pattern of handwriting data. In order to acquire the features that can represent the individuality on handwriting, Dimension Reduction (DR) approach was chosen to be applied in this study.

The capability of DR techniques in reduction the dimension of data can simplify the way to select the features. Moreover, reduction concept is capable in enhancing the data representation for classification purposed (Ji and Ye, 2009). A part from that, the selected DR technique will be injected into verification framework in order to improve the verification process. This is because the previous framework which consists of Preprocessing process, Feature Extraction process and Classification process need to be enhance in order to increase the classification accuracy. However, the verification framework that applied in this study was referring the previous framework which used by (Muda, 2009). In the same time this study proposed to observe the influences of improvement that has been done on the previous framework by additional process of Dimension Reduction. Hence, this problem statement was lead to a several research question below:

i. What is the most capable Dimension Reduction technique to be used in acquiring the individuality of handwriting?

ii. How to improve the verification framework in order to increase the accuracy of verification process?

iii. How accurate is the framework verifies the writer of handwritten document after enhancement process?

1.2 Research Objectives

Primarily aim of this study is to verify the author of handwritten document by using Writer Verification process. However, acquiring the individuality of handwriting is a major issue in this study. Furthermore, individuality of handwriting is a requirement that can influence the performance of verification process in order to verify the writer. Meanwhile, features that extracted from handwriting image will represent the individuality of handwriting. In addition, among all the features reduction process is needed to select the best feature combination to improved the verification process. Therefore, there are two objectives suggested below in order to overcome the big issue in this study:

i. To explore the Dimension Reduction Techniques for acquiring the individuality of handwriting.

ii. To enhance the framework of writer verification process by using Dimension Reduction.

iii. To evaluate the performances of verification process by using an improved framework.

The entire objective above intends to answer the research question that has mentioned in previous section. Table 1.1 below will shows the relation between research objective and the research question:

Table 1.1: Relations between Research Objective and Research Question

Research Objective	Research Question
To explore the Dimension reduction techniques for acquiring the individuality of handwriting.	What is the most capable Dimension Reduction technique to be used in acquiring the individuality of handwriting?
To enhance the framework of writer verification process by using Dimension Reduction.	How to improve the verification framework in order to increase the accuracy of verification process?
To evaluate the performances of verification process by using an improved framework.	How accurate is the framework verifies the writer of handwritten document after enhancement process?

1.3 Research Hypothesis

The hypothesis of this study is: "The proposed new frameworks by inserting Dimension Reduction techniques into previous verification framework will be able to acquire the individuality of handwriting by reducing the dimension of the features and improving the classification performance of Writer Verification."

1.4 Research Contribution

In previous verification framework, there three processes are conducted which are Preprocessing, Feature Extraction and Classification (Srihari, 2002). However, this framework still can be applied to verify the writer of handwritten document but the accuracy of verification still need to be improve. As mention before, individuality of handwriting is an important requirement in verification process to make sure the process running successfully. Therefore, this study will contribute the enhancement of the

verification framework by using Dimension Reduction technique. The purpose of enhancement the framework is to improve the accuracy of verification process into 100 % can prove the writer of the handwritten document.

As mention in Problem Statement, individuality of handwriting was become a big issues in most of handwriting analysis research field. Most of previous research was focus in solving this issue either in writer identification or writer verification (Bensefia *et al.*, 2004; Zhang *et al.*, 2003b; Zhang and Srihari, 2003a; Srihari, 2002; Zois and Anastassopoulos, 2001). Therefore, the first step will conducted in this study is exploring the DR techniques that capable in acquiring the significant or important feature that can represent the individuality of handwriting. However, the most important step is to enhance the verification framework by injecting the DR technique into the framework as one of the process before conducting the classification process. In the end of the process, the evaluation of framework performance will be conducted by using classification accuracy.

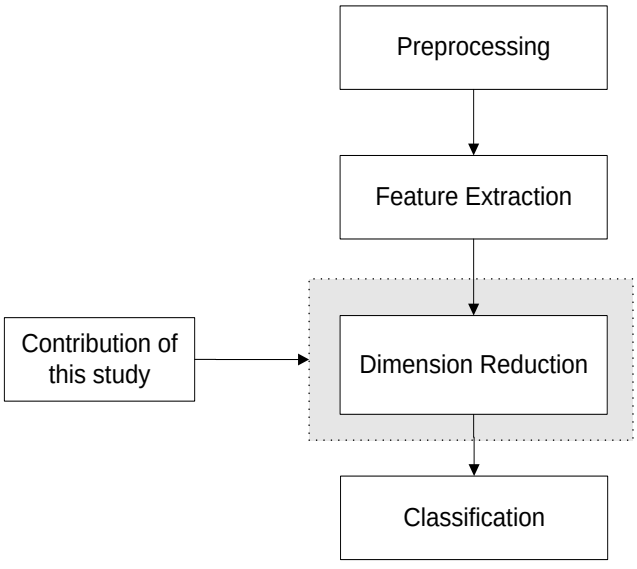

Figure 1.1: Summary of Research Contribution

1.5 Research Scope

The research scope of this study can be divided into three determinations. There are selecting the methods to formulate a framework improvement, data source and type of measurement use to compare the result. However, the major interest is development of DR Techniques and the implementation of it into previous verification framework of WV. Therefore, this study will begin by exploring the suitable methods in conducting DR task until the end of WV process which is classification task.

The first determination is to select the method to perform DR task and classification task. The purposed of DR task is to reduce the irrelevance feature without losing much information of the writer. Hence, DR approach is capable to be used in constructing the new design of WV framework. While, only one classifier will be choose after conducting some experiment for classification task to perform verification process.

The sample data is collected from IAM Handwriting Database (Marti and Bunke, 2002), where the original data that stored in the database are images. Firstly, images need to be converted into a numeric type of data before use it in the experimental. Moreover, only five writers with 3619 instances of images are chosen from 657 writers that contribute their handwriting samples. On the other hand, each writer has nine samples handwritten documents which contain more than one thousand words. Only five writers are chosen because of there is a difficulties to collect the writer has more than nine of their handwritten documents.

Finally, two types of measurement will be calculated to determine the effectiveness of verification process. First is the number of reduced features which resulted by the first objective and second is classification accuracy that resulted by second objective. The diagram below summarizes the research scope in this study:

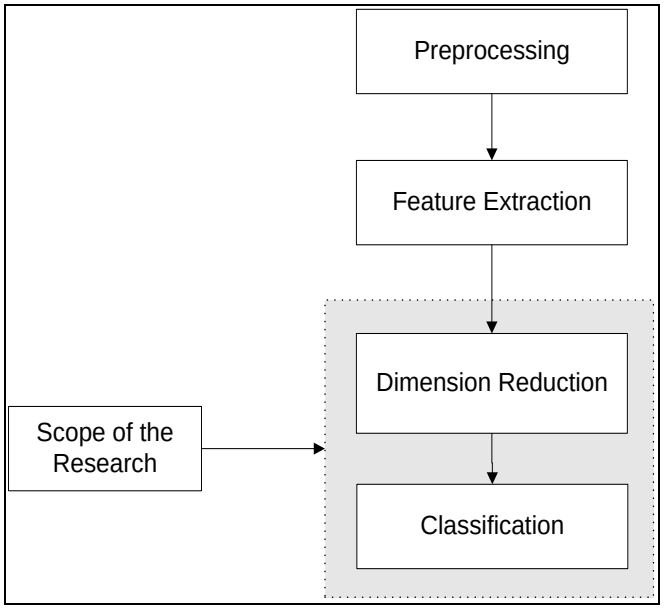

Figure 1.2: Summary of Research Scope

1.6 Research Significance

The significant of this study is depending on the effectiveness of the proposed technique to solve the main issues that has been stated in this study. This method should be able to acquire the individuality of handwriting which reflects the writer of handwritten document. The consumption of Individuality of handwriting will improve the performance of verification process and next will increase the classification accuracy. Besides that, this study will observe the influence of DR technique toward the enhancement of verification framework in achieving the objective of this study. Moreover, the individuality of handwriting that produced by selected individual feature using the proposed technique will affect an original information of the writer. Therefore, the individual features will influent the classification process in learning the training data. If the writer can 100 % be verified by using the improved verification framework, the contribution of this study in the verification algorithm is very significant to be used in a real problem that have big data. This algorithm has a wide variety of potential applications, from security, forensics, and financial activities to archeology.