CHINESE CHARACTER SEGMENTATION

TIANG KING MING

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL : CHINESE CHARACTER SEGMENTATION
SESI PENGAJIAN : 2014/2015______
Saya,____________TIANG KING MING______

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

________ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang 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: ____________________	____________________________
____________________________	Nama Penyelia
____________________________	
Tarikh: ____________________	Tarikh: ____________________

CATATAN: * Tesis dimaksudkan sebagai Laporan Projek Sarjana Muda (PSM).
** Jika tesis ini SULIT atau atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

CHINESE CHARACTER SEGMENTATION

TIANG KING MING

This report is submitted in partial fulfilment of the requirements for the
Bachelor of Computer Science (Software Engineering)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2015

DECLARATION

I hereby declare that this project report entitled

CHINESE CHARACTER SEGMENTATION

Is written by me and is my own effort and that no part has been plagiarized
Without citations.

STUDENT : _______________________________ Date: ________________
(T IANG KING MING)

SUPERVISOR : _______________________________ Date: ________________
(DR MOHD SANUSI AZMI)

DEDICATION

This is dedicated to my beloved family, thank you very much for the unconditional supports with my studies. Thank you for giving me the chance to take my desired field of studies and provide me a chance to improve myself through the journey of my life. I love you.

ACKNOWLEDGEMENT

I would like to present my thanks and appreciation to DR. Mohd Sanusi Azmi for giving assistants and advises throughout the whole semester to complete this project successfully. I have learned many things and gain a lot of knowledge under the guidance from him. Thank you for everything, DR. Mohd Sanusi Azmi.

Finally, I must acknowledge as well everyone who directly or indirectly assisted, advised and supported me on doing this Final Year Project over the semester.

ABSTRACT

This project is develop a segmentation algorithm to segment Chinese character. In order to develop an effective algorithm is necessary to study the existing segmentation algorithm/technique, including the history of segmentation technique and current segmentation technique. ut, there is the problem for segment Chinese character because of difficultly to identify the character's radical of character itself, since some Chinese character can be the radical of other Chinese character. Besides that, handwriting also a problem, because of handwritten is difficulty be recognize. Handwritten character due to different writes has their own style of writing therefore the feature of handwritten character is not constantly same, although it is written by same writer, let alone the difference writer. This project develop a segmentation algorithm using the concept of image processing like binarization and therolding to convert image into binary image. After that convert the image coordinate into horizontal and vertical histogram to identity the point for segmentation of image. After the image be segmented it will be featurize by using triangle geometric feature. Then, system will calculate the distance between the test' feature data and model data, and the distance between each character will using MAP(Mean Average Precisions) to calculate result and their ranking. In conclusion, this application consist potential for future technology. But the application still has a lot of improvement to be done in future.

ABSTRAK

Projek ini adalah untuk memcipta satu segmentasi algoritma untuk segmen cina watak. Di samping itu, belajar algoritma yang ada pada sekarang termasuk sejarah dan teknik adalah openting untuk menciptakan algoritma yang berkesan dan berhasil.Tetapi, terdapat satu masalah untul segmen cina watak kerana susah untuk mengenal pasti radikal watak yang watak itu sendiri punya, sebab sesetengah cina watak boleh dijadikan sebahai radikal cina watak lain. Selain itu, tulisan juga adalah satu masalah kerana tulisan susah untuk diiktirafkan. Mengikut kepada tulisan yang berlainan ada tulisan watak yang berlainan. Oleh itu, ciri=ciri tulisan watak juga berlainan, walaupun ia ditulis oleh penulis yang sama, apatah lagi penulis perbezaan. Projek ini mengunakan konsep pemprosesan imej atau gambar macam binarization dan therolding untuk menukarkan imej kepada binari imej untuk menciptakan satu segmentasi algotihma. Selepas itu,menukarkan imej penyelerasan kepada histogram yang melintang dan menegak untuk mengiktirafkan thet titik bagi segmentasi imej. Selepas imej disegmen, ia akan dicirikan dengan menggunakan ciri-ciri segi tiga geometri. Kemudian, sistem akan mengirakan jarak antara ujian data ciri-ciri dengan model data dan jarak antara watak dengan watak lain dengan menggunakan MAP(Mean Average Precisions) untuk mengirakan keputusan dan kedudukan mereka. Kesimpulannya, aplikasi ini mempunyai potensi bagi teknologi pada masa depan. Tetapi aplikasi ini masih ada banyak peningkatan dan tempat untuk dimajukan pada masa akan datang.

CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.0 Overview

Chapter one provides the basic ideas in the development of Chinese character segmentation application with Image processing on Java. This chapter comprises the descriptions of project background, problem statement, objective, research question, project scope, project's framework and project significance. As the ideas are well stated and structured, this chapter offers a clear and big picture to the developer about the system of the application to be developed. As Confucius said "success depends upon previous preparation, and without such preparation there is sure to be failure", the elements included in Chapter One serves as the very beginning of the preparation for this Chinese character segmentation application development. This chapter is significant in guiding and directing the developer towards the desired goals with clearly specified statements of the elements included in this chapter.

1.1 Project Background

Today, the growth of technology has lead into the development of many aspects, especially Intelligent Information Systems. Artificial intelligence (AI) technology is becoming vital for people. AI technology in our lives are closely related, for example: washing machines, Global Positioning System (GPS), smartphone, some decision making system etc. So, people nowadays do not have to complete task by themselves, but the use of Artificial intelligence. However, some task done by the experts will do better, but AI technology still important because not everyone will done the task well or even do not know how to do. For example, GPS will take you a correct route to your destination, but people who know the way may have a better route or shortcut to the destination. No matter how, GPS still give a great help to many people.

This Chinese character segmentation application was developed in order to improve the Artificial intelligence (AI) technology in segmentation technique for segmenting handwritten text. Not many people know Chinese, this application can help them to segmentation the Chinese and recognize the word. This application focus on recognize the hand writing Chinese, because nowadays many application just focus on formal Chinese word or certain type of word. This application also apply concept of image processing technique like Thresholding Binarization etc.

1.2 Problem Statements

Segmentation technique used for segmenting handwritten character a unique technique that is seldom used by people. It is also rarely to apply into a real-time application. Especially apply in Chinese handwritten text. Chinese character in formal type it has been very difficult to segmenting it, because of difficultly to identify the character's radical of character itself, since some Chinese character can be the radical of other Chinese character.

Now we are faced with a further challenge, because of handwritten is difficulty be recognize. Handwritten character due to different writes has their own style of writing therefore the feature of handwritten character is not constantly same, although it is written by same writer, let alone the difference writer. This will make the segmentation of Chinese character be more difficult, and need an effective segmentation algorithm be develop to solve this problem.

1.3 Objective

The following shows the two main objectives of this project:

i. To study segmentation technique used for segmenting handwritten character.

ii. To develop segmentation algorithm for Chinese handwritten

1.4 Project Scope

The project scope for this project as follows:

i. Chinese handwritten that contains 25 Chinese words.

ii. Focus on segmentation that based on the format given to the correspondents.

iii. Number of correspondent is 10 and number of form filled by the correspondence are 10 forms. This will contribute 10*25 = 250 words

iv. The segmented images will be stored in local drive.

v. Feature extraction Triangle Geometry Feature (Mohd Sanusi Azmi, Exploiting Features from Triangle Geometry for Digit Recognition) used to extract the segmented images.

vi. The features of segmented images are stored in the Comma Separated Value format (csv)

vii. The classification of the segmented images are not the scope of this project.

1.5 Project significance

This project exposes my ability to understand segmented happen in handwritten. This project help me:

i. understand segmentation techniques used in handwritten especially for Roman, Arabic and Chinese domain

ii. Be able to segmented Chinese Words and provide the segmented images to be processed using Triangle Geometry Feature that currently used for Roman and Arabic Digit and also for Arabic Calligraphy.

1.6 Expected Output

The expected output for this project, the Chinese words written by 10 correspondence for 10 forms that consists 25 words will be segmented into words. The number of segmented images are 250 and will be stored in the local hard drive for the images and csv for the features. This segmented images and features can be used in the classification process either unsupervised or supervised machine learning.

1.7 Conclusion

Segmentation for Chinese characters is a focus in this research. A technique to segment Chinese recognition will be develop in this research. Problem for segmenting Chinese character has been identified in this research and objective have been set in order to address the problems.

For next chapter, we will discuss the literature review and project methodology. It also describes more details about the history of segmentation technique and technique used for Roman, Arabic and Chinese character. Besides that, next chapter also include the requirement to build this application, the project schedule and milestone.

Chapter II

Literature Review and Methodology

2.1 Introduction

Chapter Two discusses about the area of study involved in this project development. All the information related to this project will be reviewed. As defined by University of Wisconsin Writing Center (2012), literature review is a "critical analysis of a segment of a published body of knowledge through summary, classification, and comparison of prior research studies, reviews of literature, and theoretical articles". Thus, many researches papers that are related to this project were studied. The review on the researches done by other parties is able to make the ideas of this project justified and provide the readers with up-to-date literature on a related topic. Besides that, developer conducted the study on the current systems, and made the comparison between the existing systems. Features of existing systems are compared with the segmentation and recognition technique also.

The methodology section describe on how Chinese character segmentation application adopts the Object oriented analysis and design (OOAD) methodology and explain the reasons for using this methodology. This chapter also include project milestone and project schedule to make sure the project will able finish on time.

2.2 Fact and findings

Handwritten text segmentation is a very difficult problem because there is a large variation in handwritten styles. For example, writer may writer the same character in different ways: shapes, size, and position of character. Even within the same word. In this project facts and findings are about the historical of segmentation technique of handwritten character especially in Roman, Arabic and Chinese character. Besides that, it also include the analysis of different technique segmentation found.

2.3 Project Methodology

The Development of Chinese character segmentation algorithm is by referring to the current existing segmentation algorithm of the Roma, Arabic, Chinese character.

The development process of Chinese character segmentation algorithm is divided into two phase: investigation phase and phase Development. Investigation phase include the previous study regarding segmentation techniques focus on three domain which are Roman, Arabic and Chinese. Development phase include a general framework build for illustrate the segmentation process for Chinese character and specific framework for describing the process of image processing method for Chinese character segmentation that include three main process :

i. . Binarization:
   a. Image input is process using Otsu technique. The noisy of the image is remove and the format of image is change from grayscale to binary scale.[Dr sanusi,phd,pg59,2011]

ii. Labelling:
   a. Binarized image will undergo the process of labeling to label “1” as the background. The background usually is white color while the image will be label “0” for showing the present of the digit. The digit is in black color.

iii. Segmentation of Chinese character

a. After the image has been binarized and label, the developing segmentation algorithm is apply in order to segmented the form (contain 25 character) into character.

Table 1 Phase of methodology

Development of Chinese character Segmentation Algorithm	
investigation phase	**Development phase**
Previous Study **1. Roman** **2. Arabic** **3. Chinese**	**Framework** **1. General Framework** • **Show the general flow of Chinese character segmentation process** **2. Specific framework** • **Describing the process segmentation in image processing method for segmentation.** • **3 main process :** **i. Binarization** **ii. Labelling** **iii. Segmentation**

2.3.1 Previous study

2.3.1.1 History of Segmentation

2.3.1.1.1 Segmentation of Roman

From the past, there are a lot of languages during the Mediterranean world were written with Phoenecian alphabet that does not contain vowel sound. In order to keep the text comprehensible, the words are kept in vertical lines or interpoints. The Greek alphabets were abandoned due to it's insignificancies and forgotten. But during the Classical period c. 200 AD, the interpuncts were continue to be used by Romans. When the whitespace delineation are introduced, the documents are drop down.

Finally, both Classical and Post-Classical Latin automatic word segmentation test is known as a useful analysis tool for now.

Pre-processing and Selection of the Dataset

Table 2 List of works included in the dataset

Author	Works
Albius Tibullus	Aliorumque Carminum
Publius Vergilius Maro	Aened, Bucolics, Edogues
Gaius Valerius Catullus	All Surviving Works
Sulpicia	Epistude
Caludius Claudianus	Panegyricus Dictus et Sextus
Sextus Propertius	Elegiae
Gaius Valerius Flaccus	Argonautica
Publius Ovidius Nason	Metamorphoses Liber I-X
Phaedrus	Fabularum Aesopiarum
Quintus Horatius Flaccus	Carmina
Publius Papinius Statius	Achilleid

The table above was formed during the c. 200 to 700 AD and formed from authors that transcribed in scriptio continua (continuous script). Because of the work of noted poets, Latin prose was delineated into lines and written on scrolls for speech. Later on, the final works from 11 noted Roman poets with around 46,000 lines were added in this dataset.

During the classical period, pre-processing text are needed and the text are capitalized, segmented and punctuated so that the modern readers able to read them well. The modern word segmentation were sustained when a simple script were added, that is labelling ‘1’ at the end of word and ‘0’ for extension but it is removed when the others are introduced.

2.3.1.1.2 Segmentation of Arabic

Arabic is written by more than 100 million people in over 20 different countries and spoken in a wide arc spread across the Middle East, North Africa and the Horn of Africa. The Arabic script derived from a type of Aramaic, with the earliest known document dating from 512 AD. The Aramaic language has fewer consonants than Arabic. The new letters were created around the 7th century by adding dots to actual letters. That's why there are several letters differing only by a single dot.

Features of Arabic writing:

i. Arabic writing is cursive which write out from right to left, rather than left to right.
ii. The Arabic letters are joined together along a writing line.
iii. Arabic contains dots and other small marks that can change the definition of a word. Often the diacritic marks representing vowels are left out, and the word must be determined from its context.
iv. The shape of the letters differ depending on whereabouts in the word they are found. The same letter at the beginning and end of the word can have a totally different appearance.

During the last 20 years, many segmentation techniques have been published to build more sturdy Arabic handwritten Character Recognition (CR) system. Based on the segmentation process, two approaches have been applied to the off-line Arabic handwritten CR system which is analytical (segmentation-based) approach and global (segmentation-free) approach. The analytical approach tries to separate the characters as in, while the global approach tries to recognize the whole meaning of the words in.

2.3.1.1.3 Segmentation of Chinese

The words such as Chinese character and Japanese character are the ideographic language that the written without any space and other word delimiters. However, the words in Chinese may compromise into few characters, sometimes until five to character for a words. Word segmentation is important to when processing the Chinese character. Sometimes, it may be has incorrect segmentation in word ambiguity and unknown word.   In addition, the accuracy of Chinese word segmentation frequency related with the performance.

In the history of word segmentation, there have segmentation ambiguity in word segmentation because the same string in a word may segmented into different words such “发展中国家” can segment into “发展中(developing)/国家(country)” or “发展(development)/中(middle)/国家(country)”. So, it may cause mismatch or incorrect when the segment Chinese word.

Previously, there has a statistical formula as contextual information formula to identifying two characters in Chinese word. Contextual information formula founded better than mutual information formula to identifying two characters in Chinese word. Contextual information formula is in the form of frequency character that adjacent to the bigram being was found that it is the factor to predict the character in Chinese word.

After that, there have another Chinese word segmentation technique that uses character position tag as the sequence labelling which learn the benefits of the segmentation configuration included the context of the character and the segmentation of previous character.