

SEGMENTATION OF JAWI HANDWRITTEN WORDS INTO SUBWORDS

MUHAMMAD NUR ILHAM BIN IBRAHIM

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS

JUDUL: SEGMENTATION OF JAWI HANDWRITTEN WORDS INTO SUBWORDS

SESI PENGAJIAN: 2014/2015

Saya MUHAMMAD NUR ILHAM BIN IBRAHIM mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) 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 dimana penyelidikan dijalankan)

TIDAK TERHAD

(TANDATANGAN PENULIS)
PENYELIA)

Alamat tetap: 105 A KG BARU
TEBING TEMBAH,
PAKA,
23100 TERENGGANU.

Tarikh: 11 Jun 2015

(TANDATANGAN

DR. MOHD SANUSI BIN AZMI
Nama Penyelia

Tarikh: 11 Jun 2015

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

SEGMENTATION OF JAWI HANDWRITTEN WORDS INTO SUBWORDS

MUHAMMAD NUR ILHAM BIN IBRAHIM

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

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2015

DECLARATION

I hereby declare that this project entitled

SEGMENTATION OF JAWI HANDWRITTEN WORDS INTO SUBWORDS

Is written by me and my own effort and no part has been plagiarized without citations.

STUDENT: _____ Date: _____

(MUHAMMAD NUR ILHAM BIN IBRAHIM)

SUPERVISOR: _____ Date: _____

(DR MOHD SANUSI BIN AZMI)

DEDICATION

To my beloved parents, Ibrahim bin Awang and Hishamunila binti Che Leh

ACKNOWLEDGEMENT

First of all, thank you for Allah s.w.t. for giving me chance and opportunity for completing this PSM project. Without His mercy, I may not have any ability or effort to do so.

Next, I want to express my gratitude to my supervisor, Dr Mohd Sanusi bin Azmi for assisting and advising along the project development. His guidance and times do give impact to make this project successful.

Lastly, I would like to express my thanks to everyone whom were involved in this project. Without those support, this project may not been completed in time.

ABSTRACT

The project is about image processing which is focusedly on words segmentation. The word segmentation mainly used in this project is Jawi or Arabic handwritten, which extracted from old Malay manuscript. The problems occur when current words segmentation algorithm which applied to Raman and Latin words cannot be directly implement into Jawi or Arabic words caused by the cursiveness and diacritics in the words. Besides, a new alternate algorithm must be develop in order to accomplish the objective of this project. The objective is to enhance the segmentation of Jawi or Arabic words segmentation, instead developing the segmentation technique to be applied on Jawi or Arabic words. The methodology used has been divided into two phase; the investigation phase which having the data extraction and investigation from past research, and also implementation phase which the coding and algorithm development has been build in four phase. The result of this project is the old Malay manuscript Jawi handwritten on it has been extracted; which is firstly converted into binary image through binarization, then segmented from words into subwords before producing the RGB image of subwords.

ABSTRAK

Projek ini adalah berkenaan pemprosesan imej dimana memfokuskan kepada segmentasi huruf. Dalam projek ini, pensegmentasian huruf dilakukan keatas tulisan Jawi ataupun Arab, dimana ianya diekstrak daripada manuskrip Melayu lama. Masalah timbul apabila algoritma pensegmentasian huruf yang wujud sekarang yang mana diaplikasi keatas tulisan Rom mahupun Latin tidak boleh diaplikasi terus keatas tulisan Jawi mahupun Arab disebabkan kursif dan diakritik-diakritik yang ada pada huruf. Selain itu, satu algoritma baru yang telah diubah perlu dibina bagi memenuhi objektif projek ini. Objektifnya adalah untuk meningkatkan pensegmentasian huruf Jawi atau Arab, selain daripada membina teknik segmentasi yang boleh diaplikasi terhadap tulisan Jawi atau Arab. Metodologi yang digunakan telah dipecahkan kepada dua; fasa penyiasatan dimana pengekstrakan data dan penyiasatan terhadap kajian lepas, serta fasa pelaksanaan dimana pengekodan dan pembinaan algoritma telah dibina dalam empat fasa. Hasil daripada projek ini ialah manuskrip Melayu lama telah diekstrak, dimana ia telah ditukar kepada imej binary melalui proses peminarian, kemudian disegmentasikan daripada sebuah perkataan kepada patah perkataan sebelum menghasilkan patah perkataan RGB.

TABLE OF CONTENTS

CHAPTER	SUBJECT	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
CHAPTER I	INTRODUCTION	
	1.1 Introduction	1
	1.2 Problem statement(s)	2
	1.3 Objective	4
	1.4 Scope	4
	1.5 Project Significance	4
	1.6 Expected Output	5
	1.7 Conclusion	5
CHAPTER II	LITERATURE REVIEW AND PROJECT METHODOLOGY	
	2.1 Introduction	6
	2.2 Facts and Findings	6
	2.2.1. Domain	10
	2.2.2. Existing System	10
	2.2.3 Technique	10
	2.3. Project Methodology	11
	2.4. Project Requirement	14
	2.4.1. Software Requirement	14

	2.4.2. Hardware Requirement	15
	2.4.3. Other Requirements	15
	2.5. Project Schedule and Milestones	16
	2.6. Conclusion	24
CHAPTER III	ANALYSIS	
	3.1. Introduction	25
	3.2. Problem Analysis	26
	3.3. Requirement analysis	27
	3.3.1. Data Requirement	27
	3.3.2. Functional Requirement	29
	3.3.3. Non-functional Requirement	30
	3.3.4. Others Requirement	30
	3.4. Conclusion	Error!
	Bookmark not defined.	
CHAPTER IV	DESIGN	
	4.1. Introduction	31
	4.2. High-Level Design	31
	4.2.1. System Architecture	32
	4.2.2. User Interface Design	33
	4.2.3. Database Design	33
	4.2.3.1. Conceptual and Logical Database Design	33
	4.3. Detailed Design	34
	4.3.1. Software Design	34
	4.3.2. Physical Database Design	39
	4.4. Conclusion	39

CHAPTER V	IMPLEMENTATION	
	5.1 Introduction	40
	5.2 Software Development Environment Setup	40
	5.3 Software Configuration Management	41
	5.3.1 Configuration Environment Setup	41
	5.3.2 Version Control Procedure	42
	5.4 Implementation Status	43
	5.5 Conclusion	44
CHAPTER VI	TESTING	
	6.1 Introduction	45
	6.2 Test Plan	45
	6.2.1 Test Organization	46
	6.2.2 Test Environment	46
	6.2.3 Test Schedule	46
	6.3 Test Strategy	47
	6.3.1 Classes of Test	48
	6.4 Test Design	48
	6.4.1 Test Description	48
	6.4.2 Test Data	49
	6.5 Test Result and Analysis	51
	6.6 Conclusion	57
CHAPTER VII	CONCLUSION	
	7.1 Observation on Weakness and Strengths	58
	7.2 Propositions for Improvement	59
	7.3 Project Contribution	59
	7.4 Conclusion	60

REFERENCES	61
BIBLIOGRAPHY	62
APPENDICES	63
Appendice 1 : Test Data	64
Appendice 2 : User Manual	66

Chapter I

Introduction

1.1. Introduction

Once, between 13th to 15th of centuries, peoples love to write the manuscript of the government at that time into the manuscript. Malay manuscript as the example, there are many practical information and the morals storied about the styles and ways of life the Malay peoples at that time.

But, most of the manuscript has been lost or terminated caused by poor written and also wrote on easily-disposed materials (leaves, woods, animal skin, etc). For the Malay manuscript, many of them has also lost because the publishing, copying, and spreading process of the manuscript happened at 19th of century, instead of there were effort from the researchers and Western orientalist.

There were many topics that has been included in those manuscript. There were some story of the Malays histories, laws, traditional herbs and medicals, language, astronomical, and many more. But, most of this manuscript has been spread through the sea out from the Malays Land (Tanah Melayu). It is said that there are only about 5000 totals of Malays manuscript currently at Malaysia meanwhile the others has been displayed or as the research item at other countries museum or universities.

Malay language has played an important role as lingua franca of trade for centuries. The diplomacy, religion, and many more aspect has used this language throughout the Southeast Asia maritime. Thus, the language then was generally

written in Jawi, a modified form of Arabic script for those purposes. *Hikayat Raja Pasai*, *Taj al-Salatin*, and *Sejarah Melayu* are the examples of famous manuscripts that has been researched about.

Meanwhile, the study on old Malay manuscript currently executed with the helps of Wan Ali Wan Mat, the master on old Malay manuscript world. During this modern era, it is more attractive to study about the manuscript, mainly wrote in Jawi handwritten, from the view of computer science. The study from this scope has been started at the end of 90's actually by Khairuddin Omar and Mazani Manaf, which then their thesis has be the main source for further research and development in Jawi words.

The effort in this study of the Jawi words which initially based on old Malay manuscript then brought to the field of image processing. In image processing, there is a technique called segmentation which applied on the words.

There are three main approaches to segment the Jawi words, which is projected histogram profile, related components labelling, and determination of the point to segment. In this report, the study of Jawi words is focusing on Jawi handwritten words at old Malay manuscript.

In general scope, word segmentation is the way of separating written language into its word components. In Jawi words which has quite similar characteristic with Arabic handwritten except the diacritics part, is differ than Roman or Latin words which the cursive give some challenge for segmenting.

1.2. Problem statement(s)

Jawi or Arabic manuscript is different than Roman or Latin words. The calligraphic words contain the diacritics; a sign, such as an accent or cedilla, which when written above or below a letter. The diacritics purposed to indicate the differences in pronunciation from the same letter when unmarked or differently marked. For examples, there are several diacritics such as "fathah", "kasrah", and "dhammah" with special spelling for each of them.

Besides, in Roman or Latin segmentation, the white spaces between the characters can be totally indicate as the separator of the words, even as subwords itself. But, difference than Jawi or Arabic, their cursiveness which some characters may cursive or slant more to other characters give a big problems and challenges to segment them. The algorithm applied to Roman or Latin words in short cannot be directly applied to Jawi or Arabic words without changing the algorithm.

The table below is the example for comparison of segmentation between Roman words and Arabic words:


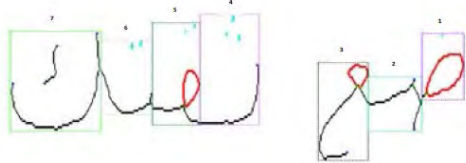
 <p>The image shows a Roman handwritten word 'CAN YOU IDENTIFY' inside a rounded rectangle. Below it, the word 'YOU' is segmented into three individual characters: 'Y', 'O', and 'U', each in its own black box.</p>	 <p>The image shows two examples of Arabic handwritten words. The first word is 'بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ' (Bismillah) with colored boxes around the letters and numbers indicating stroke order. The second word is 'بِسْمِ اللَّهِ' with similar colored boxes and stroke order numbers.</p>
<p>a. Roman handwritten</p>	<p>b. Arabic handwritten</p>

Figure 1.2 Segmentation of words for Roman handwritten and Arabic handwritten

So, there are some questions need to be answer about which are:

- i. What is the segmentation framework which is suitable and effective for calligraphic words, which can be used for segmenting the original words into subwords without even change the initial form?
- ii. What is the weakness of the current segmentation for Jawi or Arabic words?
- iii. How to proof the comparison between the original manuscript with the fake one?

1.3. Objective

The objectives of this study are listed above:

- i. To enhance the segmentation framework for segmenting Jawi or Arabic words into subwords, without changing the initial form or meaning of the words.
- ii. To develop a segmentation technique which suitable and effective to be used with Jawi or Arabic words.
- iii. To implement the features of triangle geometry formulation through geometric function.
- iv. To produce the calculation for comparison of manuscript originality.

1.4. Scope

The scope of this project as follow:

- i. A page of old Malay manuscript is taken as input.
- ii. There will be only Jawi or Arabic words in the images, excluding the diacritics.
- iii. The algorithm created will only used to segment the words into subwords and applied to Jawi or Arabic handwritten only.
- iv. Actual words cannot be recognized as there is no words library will be used.
- v. Low opacity, red blurred, indirectly cut of images used cannot be solved and may brought problems to the algorithm.
- vi. Large size of image used will take long time to extract the image as the algorithm must loop through all the words and diacritics.
- vii. The test image used must have no frame in it.

1.5. Project Significance

This project significance of this projects are:

- i. This project will help researchers to extract the words into subwords from old Malay manuscript. So, additional and deeper investigation from the results given can be process and executed.
- ii. The algorithm developed can help researchers to spend minimal time to get the subwords image. Thus, the researchers will get an efficient and reliable algorithm for segmentation purpose.
- iii. This project will catch all words contained in the image. The diacritics in the image also will be captured. This means that no single words or diacritics will be escape, making the algorithm more precised. This helps the researchers to get a detailed result.
- iv. The calculation produced will help to enhance the originality measurement. This means, the researcher will be able to have more detail data on originality better than having just an observation.

1.6. Expected Output

A set of subwords resulted from the image of Arabic manuscript will be produced based on the steps that planned from this project. The subwords, which extracted from words on the image will be produce individually, also in image form. There will be no subwords or diacritics that will be left out. Hence, the subwords can be used for further investigation.

Some result on calculation of originality comparison between original manuscript and test manuscript also will be display as the result.

1.7. Conclusion

Some overview of what the project that will be carry on has been explained at this chapter. The problem statement for the feature has been stated, as with its following objective that need to be fulfil. The scope also has been mentioned as the limit for the project.

Besides, the importance of why the project need to be executed also has been stated at the significance part. Lastly, the expected output for this project also already stated before the conclusion.

Chapter II

Literature Review and Project Methodology

2.1 Introduction

In this chapter, some literature review from many source, which mostly from the research that has been conducted and investigation made from some researchers will be discuss. There may have some review on the study of segmentation, which cover on segmentation in Arabic, Roman, and others. Plus, some comparison between this segmentation also will be discuss too.

2.2. Facts and Findings

The research of existing system is mainly about the Arabic handwritten segmentation from past research as both Jawi and Arabic words are majorly same, except for the diacritics.

There were many research has been conducted for many types and kind of words in this world. Some of the words can be applied on with same algorithm that has been made for other types of words even they are totally different. But, for Jawi

or Arabic handwritten, most of the researchers agreed that for this kind of handwritten, normal algorithm that applied on most type of words such as Roman, Latin, or even Chinese handwritten cannot be directly applied caused by the unique characteristic that have on Arabic words.

Actually, segmentation for various kind of handwritten such as Chinese, Roman, and Japanese were mostly settled, as has been stated by Abdelmalek Zidouri. Plus, there are more challenge on bad quality publishing, damage, and unlimited handwriting. Plus, there were no are no opposition on the statement that a cursive writing is quite a problem one to be segment, such as Arabic or Jawi characters. In short, a cursive writing actually is a big problems for segmentation of text to discriminate into characters.

Arabic letters has some unique characteristic itself. MS Khorshed and William F. Closksin mentioned in their research that Arabic scripts are totally cursived and very context sensitive. The connected letters has been recognizedly has four different shape that need to be focused on which are Isolated Form (IF), Beginning Form (BF), Middle Form (MF), and End Form (EF)(Abdelmalek Z.).

So, for segmenting Arabic letters, the shape of the letters are depend on its location in the word. Cursive written for Arabic also very high in form of characters variability within a word. Thus, implementing same algorithm as applied to other types of handwritten actually is failed if there are no fundamental change for current algorithm.

There were many styles and algorithm that has been made for Arabic segmentation, but there has no algorithm that actually succeed to be said that it is the best one to perform segmentation. The words, which have four forms when connected to be words each need to be distinguish. But, most of the research will detect the baseline for the words first.

Using the source of Arabic words such as from a bitmap format, the Arabic words will be differentiated. From this isolation, most of the techniques will detect the baseline; a line which will be the source or backbone for the next step of the created algorithm because the words may have contracted from, instead having two

points of words with same distribution. Besides, instead of detecting the baseline that decided from the algorithm created, there also some line that the researcher made for further purpose, such as for diacritics removal or selecting the point for isolation.

Besides, some additional step, rotating the words (generally turned into vertical form) also has been executed. This is due some researcher decide to make it more precise and details. This step also used to detect the slant angle that happen when different kind of handwriting were tested. Plus, it is also effected by the location of the words itself. This step were executed for feature extraction (Khorsheed and William).

Next, some of investigation stated that it is a need to have several approaches to extract the words, such as holistic approach and analytical approach (L Zheng et al). Holistic approach purposed to segment a word into its characters, which then will identified each of them separately. The analytical approach which require large Arabic words lexicon, purposed to identify the whole words without any effort to segment them and to recognize the primitives individually.

The recognition of character actually the structure itself before implementing the local features for complete letters. Plus, if using the Connected Component Algorithm, the pixel number for those characters will be higher as it is the structural park, meanwhile others will be set as the stress mark.

But, before that, the diacritics extraction will be made first. Diacritics, which in form of small slanted line and may also in zig-zag form need to be extracted and removed out for further recognition of the words. If not, the detection of words cannot be perform better and may affect the algorithm itself, as the result for confusing in recognizing. In short, it purposed to avoid interruption during selecting the local minimal points pieces of Arabic words localization. This process has been made using a modified version of algorithm (H. Boukerma and N. Farah).

Then, the words extracted with various techniques. M. Elzobe et al. used to have vertical projection histogram for each line of words from binary image, which then computed to produce a string indicating relative variations in pixel. The search for patterns in variation also conducted to segment the characters representatives. L.

Zheng et al. also used quite same algorithm for this purposed. M. Elzobi et al. algorithm also use the same techniques, with the solving for vertical isolated overlapping.

Meanwhile, H. Boukerma and N. Farah decide to have horizontal point estimation. The image of words initially devised in three equal partitions. From those partitions, the middle will be used as the first band of all image, before the final estimated horizontal band set as the horizontal range centered based on the feature points selected. It is detailed on which for point of words that has no feature points, they will inherit the horizontal band from the nearest point if words in the image and if there is no feature points extracted from the image, they will use the most optimal solution to baseline estimation.

Besides, Khorsheed and William has made the normalization process by using the log-polar transform to transform back the rotation into translation.

Based from various techniques, there were various kind of results has been stated out. All of those algorithms were succeed and achieved the objectives, which mostly having over than eighty five percent of successful. Even though, this does not means that the algorithms performed were the best algorithm for Arabic segmentation due to some error and weakness during the investigation.

H. Boukerma and N. Farah stated that some cases of their algorithm failed due to the errorless selection of the local minima point located at the bottom curve of small descenders. Besides, M. Elzobi et al. mentioned that their algorithm sometimes failed caused by vertical connected overlapping of the words. Over segmentation that occur also need to be dealt with in some subsequent phases.

Khorsheed and William got variation in sensitivity to feature dimensionality. But, their method will be more suitable if many more fonts are used. The recognition rate also need to be improve by using the Hidden Markov Model. Besides, even thave the good result, L. Zheng still cannot produce a hundred percent algorithm that suitable for Arabic handwritten. Furthermore, for Abdelmalek algorithm, it need to use more on Nearest Neighbour Algorithm to enhance the recognition rate. The error occur caused by the character misclassification due to the segmentation error.

As a conclusion, many algorithms and research have been made and implemented to test out for the most suitable algorithm in Arabic words segmentation. But, none of them were the best one and also in need to make further investigation and development. Many errors have been produced, but it will help more on future works with better results. Thus, it can be stated that none of the algorithms has been produced yet to make segmentation of Arabic words precise and accurate as Arabic words have large differences than other kinds of words as the cursiveness and special characteristics on them make them unique and more challenging.

2.2.1. Domain

Image processing is the domain for this project which is specifically touch on words segmentation. The words segmentation mentioned here is the segmentation of Jawi or Arabic words into subwords on old Malay manuscript. Old Malay manuscript mainly use Jawi handwritten on it with some Arabic words in some of them.

2.2.2. Existing System

There already several research on various kind of handwritten such as Roman, Latin, Jawi, and Arabic. As this project focused on Jawi or Arabic handwritten, there were several features developed on this handwritten from several researchers such as Mohammad S Khorsheed and William F Clocksin (2000), L. Zheng et al. (2004), Abdelmalek Zidouri (2007), and also Hanen Boukerma and Nadir Farah (2010), M. Elzobi et al. (2010).

Meanwhile in Malaysia, Jawi handwritten words segmentation has been developed by Khairuddin Omar (2000), Mohammad Faidzul (2010), and Mohd Sanusi Azmi (2013).

2.2.3 Technique

The technique that has been applied for this project is image processing segmentation. This technique will be used to binarize the image then segmenting the words into subwords.

2.3. Project Methodology

At this part, the research methodology that are used for this project will be discussed as the methodology is important to achieve the project objective. This project methodology contains two parts which are the investigation phase and the implementation phase. Both phases actually have their own purpose for this project.

The first phase is the investigation phase. The investigation phase is the phase where the domain research is executed.

The background study, problem statements which produce the research questions are searched and investigated. Despite to accomplish all the objectives stated, all of these tasks also needed to dig deeper to find the research scope as well as the significant to the future.

After this domain research is executed, there also needs to have some searching on the literature part. Literature part is important to know the past experiments that have been done and also to get some data which might be helpful for this segmentation research. Besides, literature review also helps in stating the level of this research.

Meanwhile, when the investigation phase is finished, this means that the implementation will be executed with the guide by the problems arising and the objective of this experiment. There are several parts at this phase which are:

- **Data collection and data analysis**

The data for this research will be collected. The data, especially on segmentation, will be collected from many sources relating to the segmentation especially the segmentation of words. This may help with the comparison of segmentation of other words styles such as Roman, Chinese,

and so on. This data then being analysed, reallocated if the data not help or not enough, and analysed again until it is satisfied.

During this part, the algorithm of the features is not being built yet, but the standard or how it must to work will be started to organized. From this phase, it has been planned that the algorithm must detect each of the Jawi or Arabic words, which are known in calligraphic forms. Besides, the algorithm also need to detect the diacritics of the Arabic font which specialised to state the sound of the same characters for Arabic words.

A binarization method for changing the actual images of any Jawi or Arabics words into binary form (1 and 0 only) is the result for this phase.

- **Segmentation of words feature development**

At this phase, the features started to be developed. From the past experiment or research, the features then being built with the guide of those. As summary, the image of calligraphic will be scanned through from y-axis to the x-axis in binarization form. Plus, it is surely the result of each line will be different as the features should be detect the words precisely. The segmentation will produce RGB image of segmented words from the binarize files.

- **Originality Comparison Calculation**

At this part, the segmented RGB image will be selected back for futher processing. The image will be choosed, either single or multiple at a time, for some calculations algorithm to check the originality with the original image

- **Result analysis.**

Based on the result which gain from the previous phase, the outcome of the features developed will be analysed. The result actually will be compare with the image that had been used to pass through the binarization and detection of bit before.