UNIVERSITI PUTRA MALAYSIA

DEVELOPMENT OF AN AUTOMATED TECHNIQUE FOR RECONSTRUCTING JAWI CHARACTERS IN HISTORICAL DOCUMENTS

TENGKU MOHD AFENDI ZULCAFFLE

ITMA 2007 2

DEVELOPMENT OF AN AUTOMATED TECHNIQUE FOR RECONSTRUCTING JAWI CHARACTERS IN HISTORICAL DOCUMENTS

TENGKU MOHD AFENDI ZULCAFFLE

MASTER OF SCIENCE
UNIVERSITI PUTRA MALAYSIA

2007

DEVELOPMENT OF AN AUTOMATED TECHNIQUE FOR RECONSTRUCTING JAWI CHARACTERS IN HISTORICAL DOCUMENTS

By

TENGKU MOHD AFENDI ZULCAFFLE

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia in Fulfilment of the Requirements for the Degree of Master of Science

March 2007

Dedicated

To

My Parents

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

DEVELOPMENT OF AN AUTOMATED TECHNIQUE FOR RECONSTRUCTING JAWI CHARACTERS IN HISTORICAL DOCUMENTS

By

TENGKU MOHD AFENDI ZULCAFFLE

March 2007

Chairman : Mohammad Hamiruce Marhaban, PhD

Institute : Institute of Advanced Technology

The old documents in Jawi script are still being used widely for references. The quality of the hard copies of those scripts will be deteriorating as time passes. Manual reconstruction may take long time if the documents are sufficiently thick. The accuracy of the document image recognition algorithms is much dependent on the level of noise on the document. Therefore, the development of the historical Jawi character reconstruction algorithm is a significant contributions to the success of the old Jawi manuscript maintenance and recognition systems.

The Background Subtraction technique has proved to be the best algorithm when historical document images were evaluated. The proposed technique has improved the algorithm by incorporating an autonomous decision making, that makes the binarization technique a scale invariant algorithm.

The prefiltering and post processing will further enhance the ability of the algorithm to remove noise from the documents. In the post binarization algorithm, separation techniques between characters with holes and without holes is introduced in order for different morphological operations to be applied to those characters. This method will enhance connection between broken characters but still preserving the originality of the document. A noise model has been developed to test the reliability of the proposed algorithm. The model was developed based on several predefined criteria. The algorithms have been implemented using Matlab software version 6.5.

The reliability of the proposed algorithms have been tested over simulated and real data. Comparison has been made between the Background Subtraction technique and the proposed method by manual inspection and mathematical evaluation. The results of the algorithms were mathematically evaluated using the Relative Foreground Area Error. Results have shown that better performance has been obtained using the proposed method. The framework managed to create historical Jawi characters more presentable. The system is not only applicable to historical Jawi characters, it can be easily adapted to any other historical characters in different languages.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Master Sains

PEMBANGUNAN TEKNIK AUTOMATIK UNTUK PEMBINAAN SEMULA AKSARA JAWI DI DALAM DOKUMEN BERSEJARAH

Oleh

TENGKU MOHD AFENDI ZULCAFFLE

Mac 2007

Pengerusi : Mohammad Hamiruce Marhaban, PhD

Institut : Institut Teknologi Maju

Dokumen lama dalam skrip Jawi masih digunakan secara meluas untuk rujukan. Kualiti salinan asal skrip tersebut akan menyusut bila masa berlalu. Pembinaan semula secara manual akan mengambil masa yang lama sekiranya dokumen tersebut adalah tebal. Ketepatan algoritma pengecaman imej dokumen adalah sangat bergantung kepada tahap hingar pada dokumen tersebut. Maka, pembangunan pembinaan semula aksara Jawi bersejarah adalah suatu sumbangan kepada kejayaan sistem penyelengaraan dan pengecaman manuskrip Jawi lama.

Teknik Penolakan Latarbelakang telah dibuktikan sebagai algoritma terbaik bila imej dokumen bersejarah dinilai. Teknik yang dicadangkan telah menambahbaik algoritma tersebut dengan menyelitkan pembuat keputusan secara automatik menyebabkan teknik perduaan suatu algoritma tidak ubah

skala. Pra penapis dan pemprosesan susulan akan menambah tingkatkan keupayaan algoritma untuk menyingkirkan hingar daripada dokumen. Dalam algoritma perduaan susulan, teknik pemisahan antara aksara-aksara dengan lubang and tiada lubang diperkenalkan supaya operasi morfoloji berbeza digunakan untuk aksara-aksara tersebut. Kaedah ini akan menambahbaikkan sambungan antara aksara-aksara yang pecah tetapi masih mengekalkan keaslian dokumen. Model hingar telah dibangunkan untuk menguji kebolehupayaan algoritma yang dicadangkan. Program-program telah dibangunkan menggunakan perisian Matlab versi 6.5 sebagai bahasa pengaturcaraan.

Kebolehupayaan algoritma yang dicadangkan telah diuji ke atas data-data simulasi dan sebenar. Perbandingan telah dibuat antara teknik Penolakan Latarbelakang dan kaedah yang dicadangkan dengan pemeriksaan manual dan penilaian secara matematik. Hasil algoritma tersebut dinilai secara matematik menggunakan Ralat Kawasan Latarhadapan Relatif. Hasilnya menunjukkan prestasi yang lebih baik telah diperolehi dengan menggunakan kaedah yang dicadangkan. Rangka kerja ini telah berjaya membuatkan character Jawi bersejarah lebih baik rupanya. Sistem ini tidak hanya boleh diaplikasikan kepada aksara Jawi bersejarah, ianya mudah diubahsuai kepada sebarang aksara bersejarah dalam bahasa yang lain.

ACKNOWLEDGEMENTS

First, all praise to almighty ALLAH SWT. The only creator, sustainer and efficient assembler of the world, for giving me strength, ability, and patience to complete this research.

I would like to express my appreciation to the Chairman of my supervisory committee Dr. Mohammad Hamiruce Marhaban. I would like to thank him for his cheery nature, assistance, and mentorship throughout the years. His patience and guidance have been invaluable to me. Without the time he has invested brainstorming ideas and discussing results, this research would never have gotten off the ground.

I would like to acknowledge and thank to Assoc. Prof. Dr. Abdul Rahman Ramli, for accepting me as one of his postgraduate students. He had patiently read through my thesis and technical papers. This dissertation would have never been completed without his help. However this is not the most important thing he has done to me. Usually, I see him demonstrating a love for teaching and supporting his students. I count myself fortunate to have known and worked with Assoc. Prof. Dr. Abdul Rahman Ramli.

In addition to thanking my members of supervisory committee, I would like to thank to the staff members of Institute of Advanced Technology especially to Mrs. Rosiah Osman for their assistance.

I would like to express my sincere gratitude to Managing Director of German-Malaysian Institute, Mr Yusof Sahir, Head of Department Industrial Electronics, Miss Jamilah Ali, Head of Section Computer and Communication Technology , Mr. Zulkifli Naim for allowing me to pursue my master degree through part time study and also for their supports through out my study.

I would like to deeply thank my parents, brothers and sisters and my friends for their unwavering support, best wishes, and encouragement through good and bad times.

I certify that an Examination Committee has met on 8th March 2007 to conduct the final examination of Tengku Mohd Afendi Zulcaffle on his Master of Science thesis entitled "Automated Historical Jawi Characters Reconstruction Technique" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Samsul Bahari, PhD
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Shattri Mansor, PhD
Professor
Faculty of Engineering
Universiti Putra Malaysia
(Internal Examiner)

Mohamed Othman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Dzulkifli Mohamad, PhD
Associate Professor
Faculty of Computer Graphic and Multimedia
Universiti Teknologi Malaysia
(External Examiner)

HASANAH MOHD GHAZALI, PhD
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee are as follows:

Mohammad Hamiruce Marhaban, PhD
Lecturer
Faculty of Engineering,
Universiti Putra Malaysia
(Chairman)

Abdul Rahman Ramli, PhD
Associate Professor
Institute of Advanced Technology
Universiti Putra Malaysia
(Member)

_____
AINI IDERIS, PhD
Professor/Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 14 JUNE 2007

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

_____

TENGKU MOHD AFENDI ZULCAFFLE

Date: 23 APRIL 2007

TABLE OF CONTENTS

			Page
DEDICATION			ii
ABSTRACT			iii
ABSTRAK			v
ACKNOWLEDGEMENTS			vii
APPROVAL			ix
DECLARATION			xi
LIST OF TABLES			xiv
LIST OF FIGURES			xv
LIST OF ABBREVIATIONS			xxi
CHAPTER			
1	**INTRODUCTION**		1
	1.1	Background	1
	1.2	Problem Statement	3
	1.3	Motivation	4
	1.4	Objectives	5
	1.5	Scope of Thesis	5
	1.6	Research Contribution	6
	1.7	Thesis Organization	7
2	**LITERATURE REVIEW**		8
	2.1	Document Image Analysis and Understanding in Jawi Script	8
	2.2	Binarization Method	9
		2.2.1 Global Thresholding	9
		2.2.2 Local Thresholding	21
		2.2.3 Evaluation of Binarization Method	39
	2.3	Euler Number	43
	2.4	Edge Detection	46
	2.5	Prefiltering of Document Images for Binarization	53
	2.6	Binary Morphological Dilation in Document Reconstruction	55
	2.7	Summary	57
3	**METHODOLOGY**		59
	3.1	Introduction	59
	3.2	Historical Jawi Text Region Noise Modeling	60
	3.3	Automated Historical Jawi Character Reconstruction Technique	66
		3.3.1 Prefiltering	64
		3.3.2 Thresholding Algorithms	65
		3.3.3 Post Processing	74
	3.4	Summary	81

4	**RESULTS AND DISCUSSION**		83
	4.1	Introduction	83
	4.2	Historical Jawi Text Region Noise Model	84
	4.3	Automated Historical Jawi Character Reconstruction Technique	92
	4.4	Comparative Study between Background Subtraction Technique and Automated Historical Jawi Character Reconstruction Technique	105
	4.5	Summary	118
5	**CONCLUSION AND RECOMMENDATION**		120
	5.1	Conclusion	120
	5.2	Suggestions for Future Works	122
REFERENCES			124
APPENDIX			136
BIODATA OF THE AUTHOR			149
LIST OF PUBLICATIONS			150

LIST OF TABLES

Table		Page
4.1	The initial threshold values of a few samples of the simulated and real data for the Iterative Thresholding.	96
4.2	The results of the real data for the BS and the PITBR using RAE	110
4.3	The results of the real data for the BS and AHJCRT using RAE	115

LIST OF FIGURES

Figure		Page
1.1	The Jawi script	02
2.1	Three pixel classes in the Integral Ratio thresholding approach	19
2.2	Local windows definition	34
2.3	Roberts Cross convolution masks	46
2.4	Pseudo-Convolution masks used to quickly compute approximate gradient magnitude	47
2.5	Sobel convolution masks	48
2.6	Prewitt convolution masks	48
2.7	Pseudo-convolution masks used to quickly compute approximate gradient magnitude	48
3.1	Degradation model of historical Jawi document	62
3.2	The AHJCRT flowchart	63
3.3	The proposed thresholding algorithm	69
3.4	The flowchart of the post processing of AHJCRT	74
3.5	Structuring Element of S_{DH}	80
4.1	A sample of 12 handwritten Arabic word taken from IFN/ENIT-database.	82
4.2	The point spread function of the mLoG with σ = 0.1	82
4.3	The complemented image of Figure 4.1	82
4.4	The input image after filtered by mLoG with σ values 0.1, 0.2, and 0.3	83
4.5	The point spread function of the mLoG with σ = 0.4	84

4.6	The input image after filtered by mLoG with σ value = 0.4	84
4.7	The point spread function of mLoG with σ = 0.5	84
4.8	The input image after filtered by mLoG with σ value = 0.5	84
4.9	The grayed background image with the σ of mLoG = 0.1, 0.2 and 0.3	85
4.10	The grayed background image with the σ of mLoG = 0.5	85
4.11	The point spread function of motion blurring with *len* = 10 and $\theta = 0$	85
4.12	The point spread function of motion blurring with *len* = 15 and $\theta = 45$	85
4.13	The point spread function of motion blurring with *len* = 15 and $\theta = 90$	86
4.14	A sample of the pixel values taken from the image after background graying with σ of mLoG = 0.1, 0.2 and 0.3	87
4.15	Result of motion blurring with *len* = 10 and $\theta = 0$ to the pixels of the Figure 4.14	88
4.16	A sample of the pixel values taken from the image after background graying with σ of mLoG = 0.5	88
4.17	The output of motion blurring with *len* = 10, $\theta = 45$ with the input taken from the image in Figure 4.16	89
4.18	The final output of HJTRNM with parameters of motion blurring with *len* = 10 , $\theta = 0$ and σ of mLoG = 0.1, 0.2 and 0.3	89
4.19	The final output of HINM with parameters of motion blurring with *len* = 15, $\theta = 45$ and σ of mLoG = 0.5	89
4.20	The final output of HJTRNM with parameters of motion blurring with *len* = 15, $\theta = 90$ and σ of mLoG = 0.7	90
4.21	The output of Wiener to the HJTRNM with parameter of *len* = 10, $\theta = 0$ and σ of mLoG = 0.1, 0.2 and 0.3	90
4.22	The output of Wiener to the HJTRNM with parameter of *len* = 15, $\theta = 45$ and σ of mLoG = 0.5	90

4.23	The output of Wiener to the HJTRNM with parameter of *len* = 15, θ = 90 and σ of mLoG = 0.7	91
4.24	Real data 1	91
4.25	Real data 2	91
4.26	Real data in Figure 4.24 filtered by 3x3 Wiener	92
4.27	Real data in Figure 4.25 after filtered by 3x3 Wiener	92
4.28	The intensity different between Figure 4.21 and its estimated background	93
4.29	The intensity different between Figure 4.22 and its estimated background	93
4.30	The intensity different between Figure 4.23 and its estimated background	93
4.31	The intensity different between Figure 4.26 and its estimated background	93
4.32	The intensity different between Figure 4.27 and its estimated background	94
4.33	The result of Figure 4.28 after binarized by the Iterative Thresholding	94
4.34	The result of Figure 4.29 after binarized by the Iterative Thresholding	95
4.35	The result of Figure 4.30 after binarized by the Iterative Thresholding	95
4.36	The result of Figure 4.31 after binarized by the Iterative Thresholding	95
4.37	The result of Figure 4.32 after binarized by the Iterative Thresholding	96
4.38	A long connected letter that consists of two isolated letters	96
4.39	One part of the long connected letter in Figure 4.38 that categorize into characters without holes	97
4.40	Another part of the long connected letter in Figure 4.38 that categorize into characters with holes	97

4.41	Result of morphological operations and combination process of Figures 4.39 and 4.40	97
4.42	Simulated image for long connected letters that have thin connection	98
4.43	One part of the long connected letter in Figure 4.42 that categorize into characters without holes	98
4.44	Another part of the long connected letter in Figure 4.42 that categorize into characters with holes	98
4.45	Result of morphological operations and combination process of Figures 4.43 and 4.44	98
4.46	The character that perceived as character with hole by Euler Number method	99
4.47	The subsequent effect of Euler Number method after the thickening process	99
4.48	The subsequent effect of the new holes detection method after the dilation process	100
4.49	Simulated image that has a similar defect as in Figure 4.46	100
4.50	The subsequent effect of Euler Number method after the thickening process	100
4.51	The subsequent effect of the new holes detection method after the dilation process	100
4.52	A long connected character with one hole before separation process	101
4.53	The character in Figure 4.52 after dilation and thinning processes when separation technique was not applied to the system	101
4.54	The character in Figure 4.52 after dilation and thinning processes when separation technique was introduced to the system	101
4.55	The character in Figure 4.42 after dilation and thinning processes when separation technique was not applied to the system	102
4.56	The character in Figure 4.42 after dilation and thinning processes when separation technique was introduced to the system	102

4.57	A real data before applying spurious effect removal technique	102
4.58	A simulated data before applying spurious effect removal technique	102
4.59	Figure 4.57 after applying spurious effect removal technique	102
4.60	Figure 4.58 after applying spurious effect removal technique	103
4.61	The result of BS technique over the document image in Figure 4.18	103
4.62	The result of the ITBR technique over the document image in Figure 4.18	103
4.63	Image in Figure 4.25 after binarized by BS	104
4.64	Image in Figure 4.25 after binarized by ITBR	105
4.65	The result of the PITBR technique over the document image in Figure 4.18	106
4.66	The result of the PITBR technique over the document image in Figure 4.25	106
4.67	Error of the BS and PITBR with *len* = 10 and $\theta = 0$	107
4.68	Error of the BS and PITBR with *len* = 10 and $\theta = 45$	108
4.69	Error of the BS and PITBR with *len* = 10 and $\theta = 90$	108
4.70	Error of the BS and PITBR with *len* = 15 and $\theta = 0$	109
4.71	Error of the BS and PITBR with *len* = 15 and $\theta = 45$	109
4.72	Error of the BS and PITBR with *len* = 15 and $\theta = 90$	110
4.73	Manually constructed ground truth of Figure 4.25	111
4.74	The result of the AHJCRT technique over the document image in Figure 4.18	112
4.75	The result of the AHJCRT technique over the document image in Figure 4.25	112
4.76	Error of the BS and AHJCRT with *len* = 10 and $\theta = 0$	113
4.77	Error of the BS and AHJCRT with *len* = 10 and $\theta = 45$	113

4.78 Error of the BS and AHJCRT with *len* = 10 and $\theta = 90$ 114

4.79 Error of the BS and AHJCRT with *len* = 15 and $\theta = 0$ 114

4.80 Error of the BS and AHJCRT with *len* = 15 and $\theta = 45$ 115

4.81 Error of the BS and AHJCRT with *len* = 15 and $\theta = 90$ 115

LIST OF ABBREVIATIONS

HADMA	Hadamard Mutliresolution Analysis
NIR	Native Integral Ratio
QIR	Quadratic Integral Ratio
DTA	Dynamic Thresholding Algorithm
IFA	Integrated Function Algorithm
DoG	Differential of Gaussian
ROIs	Regions of Interest
LoG	Laplacian of Gaussian
HJTRNM	Historical Jawi Text Region Noise Model
AHJCRT	Automated Historical Jawi Character Reconstruction Technique
MLoG	Matlab version of Laplacian of Gaussian
BS	Background Subtraction
ITBR	Iterative Thresholding with Background Removal
PITBR	Prefiltered Iterative Thresholding with Background Removal
RAE	Relative Foreground Area Error

CHAPTER 1

INTRODUCTION

1.1 Background

Over the past decades, with the advancements of computer technology, digital computational techniques, and image processing technology which deals with one of the major information sources of human being, have experienced tremendous development. Availability of electronic imaging tools and effective image processing makes it feasible to enhance degraded images. Many algorithms have been developed to improve degraded historical document images. Those algorithms can be categorized into parametric and nonparametric.

The Jawi script is an art of writing that has been existed for centuries in the South East Asia. Its existence is directly related to the dawn of Islam to the South East Asia. The Jawi script is originated from Arabic script. The script has been adapted to suit Malay Language system. The Jawi alphabets and its differences with the Arabic script is shown in Figure 1.1. The figure shows the Jawi script with the circled alphabets are unique to Malay language and cannot be found in Arabic script. The Jawi script had been widely used in every aspect of life since the age of Pasai Islamic Government and later to the age of Malacca Empire and also to the age of Acheh Government on the 17^{th} century. The prove of Jawi script existence in Malaysia is when the Batu Bersurat Terengganu was found dated 702H or 1303AD (Hashim,