Faculty of Information and Communication Technology

SEGMENTATION OF TWO TOUCHING HANDWRITTEN ARABIC CHARACTERS USING OVERLAPPING SET THEORY AND GRADIENT ORIENTATION

Inam Ullah

Doctor of Philosophy

2020

SEGMENTATION OF TWO TOUCHING HANDWRITTEN ARABIC CHARACTERS USING OVERLAPPING SET THEORY AND GRADIENT ORIENTATION

INAM ULLAH

A thesis submitted
in fulfillment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2020

DECLARATION

I declare that this thesis entitled "Segmentation of Two Touching Handwritten Arabic Characters Using Overlapping Set Theory and Gradient Orientation" 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 : Inam Ullah

Date : ……..

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Doctor of Philosophy.

Signature :…….....

Supervisor Name : Associate Professor Ts. Dr. Mohd Sanusi bin Azmi

Date :…….....

DEDICATION

I would like to dedicate this research to my entire family members, where I should, in prior mention my parents, particularly my mother (late), whose prayers and support enabled me to carry out this given task. Besides, I should also mention with dedication all my siblings who, in every step of this task, instilled their spirited passion into my consideration while accomplishing this very task.

ABSTRACT

Image segmentation of offline Arabic handwritten documents is an active research area but requires efforts to segment image into regions compared to human vision, especially for degraded handwritten historical documents. Therefore, these valuable degraded handwritten documents attract researchers from all around the world but facing problems in segmentation of Arabic text because of overlapping and touching character. The overlapping and touching of character occurs by not following the standard rule of writing where, two or more characters share the same space and these touching characters are considered as one sub-word. At present many techniques are available for touching handwritten character segmentation by using the concept of connected components. These methods are easy to implement and provide high accuracy in some cases but they fail in many cases because some manual decision value is required to determine the correct segmentation path near junction point, which produce unstable character boundary. Besides, these methods are unstable when applied to handwritten characters having loops or circular path in both touching characters. In this case, the cut-point is located in incorrect place, which can lead to incorrect dividing path of a character boundary. The selection of path near junction point is one of the main challenge in segmentation of connected components. Currently, these methods contain many disadvantages usually implemented for only one layout and fonts types because of variation in writing. Apart from connected components methods, template based segmentation is another available method where several studies have been developed based on template creation for touching characters. The disadvantage is creating many templates for all possible touching types. Therefore, due to variation in writing connected components methods still unexplored especially for the cursive based handwriting like Arabic and Jawi. In this work, three objectives are highlighted, first is to identify junction point of touching image, second is to formulate direction near junction point and third is for segmentation of touching characters. The research methodology consists of three proposed ideas: junction point detection, formulate direction and segmentation stage. In junction point identification stage overlapping set theory is used to identify the segmentation point of the two touching characters. In formulate direction stage; gradient technique is used to formulate the right direction near junction point. In segmentation stage contour tracing technique is used to segment the two touching character into isolated characters. The three proposed methods were tested on IFN/ENIT, AHDB and IAM datasets. Experiments were conducted on finding of junction point where success rate is 93.3%, for the second proposed method, the success rate is 98% and last proposed segmentation method is 97.27%. In conclusion, the proposed segmentation method outperforms the existing research in term of accuracy. Proposed methods do not use any recognizer or template to control segmentation accuracy. Finally, the proposed segmentation method was again compared with state of the art methods, and it also gained better accuracy rate for degraded, non-degraded document images and the accuracy for the overall processes for AHDB is about 97.45% and 85.03% for IAM dataset.

PENEMBERENGAN DUA HURUF ARAB BERSENTUH MENGGUNAKAN SET TEORI PERTINDANAN DAN ORIENTASI KECERUNAN

ABSTRAK

Dokumen tulisan tangan Arab luar talian adalah area aktif kajian tetapi memerlukan usaha untuk penemberengan imej kepada bahagian dibandingkan dengan visual manusia terutama pada imej terhakis dokumen sejarah. Maka, dokumen terhakis yang bernilai menarik minat penyelidik dari serata dunia tetapi menghadapi masalah-masalah penemberengan terutama pada teks Arab kerana tindanan dan sentuhan huruf. Tindanan dan sentuhan tidak mengikut piawai penulisan dimana dua atau tiga huruf berkongsi ruang yang sama dan dianggap sebagai sub-perkataan. Pada masa sekarang banyak teknik yang ada untuk penemberengan sentuhan huruf tulisan tangan dengan menggunakan konsep komponen-komponen terhubung. Kaedah-kaedah ini adalah mudah untuk diimplementasi dan memberikan ketepatan yang tinggi pada keadaan tertentu tetapi gagal di dalam banyak keadaan kerana parameter penemberengan dilakukan secara manual pada titik laluan terhampir yang menghasilkan sempadan huruf tidak stabil. Selain itu, Kaedah-kaedah ini tidak stabil apabila diaplikasi kepada tulisan tangan yang mempunyai lingkaran atau bulatan pada huruf tersentuh. Untuk situasi ini, titik potongan dikenalpasti adalah tidak tetap yang membawa kepada laluan potongan yang tidak betul pada sempadan huruf. Pemilihan titik laluan terhampir adalah cabaran yang terbesar dalam penemberengan komponen terhubung. Pada masa ini, kaedah-kaedah ini mengandungi banyak kelemahan kebiasaannya pada latar dan jenis huruf kerana variasi tulisan. Sebahagian dari kaedah komponen terhubung, penemberengan berasaskan acuan adalah kaedah yang ada berasaskan kepada kajian berasaskan kepada pembentukan acuan untuk huruf bersentuh. Keburukannya perlu mencipta banyak acuan untuk semua kemungkinan huruf berhubung. Maka, hasil dari variasi di dalam penulisan komponen terhubung kaedah ini masih tidak dieksplorasi terutama penulisan berasaskan tulisan kursif seperti tulisan Arab dan Jawi. Tiga objektif dikemukakan, pertama mencadangkan kaedah untuk mendapatkan titik simpang antara huruf. Kedua, berdasarkan kepada lengkuk tulisan yang tepat berhampiran dengan titik simpang, dan ketiga ialah menembereng huruf-huruf bersentuh. Kajian metodologi merangkumi tiga idea: mengenalpasti titik simpang, formulasi arah dan penemberengan. Fasa titik simpang set teori tindanan digunankan untuk mengenalpasti titik penemberengan dua huruf bersentuh. Pada fasa formulasi arah, teknik kecerunan digunakan untuk mendapatkan arah berhampiran yang betul. Pada fasa penemberengan, teknik penjejakan kontor digunakan untuk menembereng huruf bersentuh. Ketiga-tiga kaedah yang dicadangkan telah diuji menggunakan dataset IFN/ENIT, AHDB dan IAM. Eksperimen telah dilaksanakan untuk mendapatkan titik simpang yang mana kejayaan ialah 93.3%. Untuk kaedah kedua, kejayaan ialah 98% dan yang terakhir adalah 97.27%. Kesimpulannya, metod yang dicadangkan lebih baik dari metod sedia ada dari segi ketepatan. Metod yang dicadangkan tidak menggunakan sebarangan pengecam atau acuan untuk mengawal ketepatan penemberengan. Akhirnya, metod yang dicadangkan adalah lebih baik dari metod sedia ada dan memberikan ketepatan yang lebih tinggi untuk imej yang baik dan terhakis. Ketepatan untuk dataset AHDB ialah 97.45% dan 85.03% untuk dataset IAM.

ACKNOWLEDGEMENTS

Allah, The Almighty, is the only one deserving of all praise and glory, for He is the only One who can set our paths straight. Mohammad (P.B.U.H), The Last Prophet of Allah Almighty is the one who shows mankind the path of enlightenment.

Research is a tedious task which cannot be completed in isolation. I would like to express my special thanks and appreciation to my supervisor Associate Professor Ts. Dr. Mohd Sanusi bin Azmi. His advice on my research as well as on my career has been invaluable. His support and encouragement allow me to grow as research scientist.

Special thanks to my family. I am short of words to express my feelings that how grateful I am to my mother (late), father, brothers and sisters for all the sacrifices that they made in this behalf. Also, I would like to express my appreciation to my beloved wife, son Mohammad Khan Sahibzada and daughters Fatima Inam , Zainb Inam and Maryam Inam for spending sleepless nights during my study.

The support and help of friends and the guidance of teachers is of utmost important in the process. I express my deepest gratitude to those who helped me in this journey.

I would like to dedicate this research work to my late mother and late sister. May Allah give them highest rank in Jannah Ameen Ya Rab-bul-Alameen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ANN	_	Artificial Neural Network
CCs	_	Connected Components
HMM	_	Hidden Markov Model
JUDOCA	_	Junction Detection Operator based on Circumferential Anchor
KHATT	_	KFUPM Handwritten Arabic Text
OCR	_	Optical Character Recognition
PAW	_	Piece of Arabic Word
RGB	_	Red Green Blue
SVM	_	Support Vector Machine
TC	_	Touching Character
TPS	_	Thin Plate Splines

LIST OF PUBLICATIONS

Inam Ullah, Azmi, M. S., Ishak D. M., and Yazan, M., 2019. Segmentation of Touching Arabic Characters in Handwritten Documents by Overlapping Set Theory and Contour Tracing. *International Journal of Advanced Computer Science and Applications (IJACE)*, 10(5), pp. 155-160.

Inam Ullah, Azmi, M. S., and Ishak, D. M., 2019. Junction Point Detection and Identification of Broken Character in Touching Arabic Handwritten Text using Overlapping Set Theory. *International Journal of Advanced Computer Science and Applications (IJACE)*, 10(6), pp. 256-260.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Segmentation of offline handwritten touching Arabic characters, which is an active research area but facing many challenges because of variation in writing, overlapping and touching characters. When two characters are connected with each other, the segmentation becomes difficult (Giuseppe et. al., 2017). A proper identification of connections among characters is crucial for subsequent character segmentation, since a wrong segmentation decreases the accuracy of pattern recognition algorithms. Thus, segmentation of touching characters is fundamental for OCR systems that aim to achieve good recognition accuracy. Therefore, this research focused on the segmentation methods for overlapping of two handwritten Arabic characters. Arabic is the official language in many countries (Black et. al., 2006) and writing text in Arabic language is cursive in nature and direction of writing is from right to left compared to English language, in which writing is from left to right both in printed and handwritten documents (Elgammal and Ismail, 2001; Khorsheed, 2002; Sari et. al., 2002; Abdullah et. al., 2012; Alkhateeb, 2015). Also, there is a lot of variation in writing of Arabic text (Saba et. al., 2013; Bahashwan et. al., 2014). Therefore, due to variation in writing there are great possibilities that many characters can overlap with other character in same word or other word in different line (Gaur et. al., 2014). As handwritten character segmentation is facing many challenges but in the presence of overlapping character multiplies the segmentation of touching characters.

Due to highlighted problems in presence of overlapping characters, the machine

can’t recognize the touching characters directly because there is a gap between human and machine abilities in reading handwriting text under noisy conditions especially for overlapped Arabic documents (Alaei et. al., 2018). But still libraries and national archives have huge amount of historical documents in paper format that have not been converted into computer understandable format (Drira, 2006). Although a lot of work has been done in converting these historical documents into machine readable format for better access and for further research but still these valuable data are in the form of scanned images. It means that these important and valuable documents are digitized and saved only in imaged format, which is not sufficient to provide all information (Dey et. al., 2013; Aouadi et. al., 2014) because scanning of these large volumes of paper documents and save it in image format is a very tedious work. But it is also true that most of these precious documents are degraded and need proper attention to store them in electronic format because these historical and old manuscripts have variety of interest, and could be used to study development of writing style and techniques of writing by passage of time. This gives understanding about these historical documents that how these historical and cultural changes have huge influenced on the society. It also gives knowledge about every character, ligatures, punctuations and abbreviations and, the way they have evolved enables the paleographers and historians identify the periods in which a manuscript was written. The quantity of these ancient manuscripts stored in archives, libraries and private collections is enormous and it will be useful to develop computer systems that could subsequently help the paleographers in manuscripts dating, classification and authentication (Descartes and Imran, 2009).

Thus by exploring literature, many algorithms are available for solving different types of problems especially in offline handwritten documents but still there exists unsolved problem in character segmentation of overlapping handwritten Arabic characters. The research done for overlapping Arabic characters are connected component methods

and template based segmentation methods. Connected component methods used morphological operations for selection of character boundary near junction point. While template based methods depend on the template stored in a dictionary file. The problem in these existing researches are selection of correct template, any incorrect template selection can affect segmentation accuracy because of variation in writing (Addakiri et. al., 2012). Therefore, these methods are not sufficient because they cannot provide complete information about character boundary in touching character images.

As, segmentation is considered as very important and key step, which divides an image into sub-units such as Line, words and characters, segmentation is directly related to recognition step (Lawgali, 2015). In simple words, to separate objects of interest from background. There exists many segmentation algorithms in the literature, which handle single language documents in an efficient way but can't apply to other layout or language, which is the main drawback of the available algorithms (Echi et. al., 2014; Eskenazi et. al., 2017; Alaei et. al., 2018). But still needs efforts to develop, modify or expand these available techniques to use it for multiple language segmentation (Honarpisheh et. al., 1999; Elgammal and Ismail, 2001; Loret and Palomar, 2011). For this purpose, made literature survey for this research and selected some existing segmentation methods, finally focus on segmentation of overlapping characters. Especially work related to handwritten Arabic calligraphy.

Thus, this chapter is organized as follows: Section 1.2; consists of research background of this study, which is "Segmentation of handwritten Arabic overlapping characters". Its main purpose is to highlight the problems and importance of this research especially for historical documents and text written in Arabic calligraphy. Section 1.3; is for the problem statements, which mainly discussed the problem of existing segmentation methods. These problems are highlighted in this section. In Section 1.4 research question

is identified, while Section 1.5 highlighted objectives of this study, Section 1.6 research significance is explained, research scope in explained in Section 1.7, limitation of the research in Section 1.8, Section 1.9 includes expected outcomes from this research and Section 1.10 is for thesis organization.

1.2 Research background

Based on the discussion in introduction section, several methods for performing segmentation of touching characters have been developed in the recent years. These methods for character segmentation are vertical projections, pitch estimation or character size, contour analysis, or segmentation and recognition coupled techniques (Lu, 1995; Lu and Shridhar, 1996). However, the state of the art does not provide a comprehensive answer to the problem. As a result, there is no standard approach for the segmentation of touching handwritten Arabic characters. As, in handwritten historical and degraded documents characters are touching with each other in such a way that cannot separate into character easily. Therefore, facing a lot of problem in segmentation of these Arabic historical documents into two or characters because these characters share the space (Aouadi et. al., 2016).

Focus on this research is on segmentation of two overlapping characters in these historical documents and overlapping characters of handwritten Holy Al-Quran. Segmentation process for segmenting Al-Quran needs to be studied carefully (Galas, 1998) because Al-Quran is the Holy Book and any incorrect segmentation will affect the Holiness of this Sacred Book (Melhem et. al., 2017). Segmentation can be defined as dividing given image into smaller similar regions according to some condition and eliminating unnecessary information (Khan et. al., 2014; Marmanis et. al., 2018).

Image segmentation, which is a very important step, divides the image into smaller