UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

ILLUMINATION REMOVAL AND TEXT SEGMENTATION FOR AL-QURAN USING BINARY REPRESENTATION

Laith Nazeeh Jamil Bany Melhem

Master of Computer Science (Internetworking Technology)

2015

ILLUMINATION REMOVAL AND TEXT SEGMENTATION FOR AL-QURAN USING BINARY REPRESENTATION

LAITH NAZEEH JAMIL BANY MELHEM

A thesis submitted
in fulfilment of the requirements for the degree of Master of Computer Science
(Internetworking Technology)

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2015

DECLARATION

I declare that this project entitled “Illumination Removal and Text Segmentation for Al-Quran Using Binary Representation” is the result of my own research except as cited in the references. The project has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....

Name : LAITH NAZEEH JAMIL BANY MELHEM

Date : ....

APPROVAL

I hereby declare that I have read this project and in my opinion this project is sufficient in terms of scope and quality for the award of Master of Computer Science (Internetworking Technology).

Signature : ....

Supervisor Name : Dr. MOHD SANUSI AZMI

Date : ....

DEDICATION

I would like to present my work to those who did not stop their daily support since I was born, my dear mother and my kindness father. They never hesitate to provide me all the facilities to push me foreword as much as they can. This work is a simple and humble reply to their much goodness I have taken over during that time. Also to my brothers (Mohammad, Hamzah), sisters (Rawan, Rana, Shoroq), my Grandfather, my Grandmother, my Aunt, my Uncle, friends and those entire how I love (Allah's bless all of them).

ABSTRACT

Segmentation process for segmenting Al-Quran needs to be studied carefully. This is because Al-Quran is the book of Allah swt. Any incorrect segmentation will affect the holiness of Al-Quran. A major difficulty is the appearance of illumination around text areas as well as of noisy black stripes. In this study, we propose a novel algorithm for detecting the illumination on Al-Quran page. Our aim is to segment Al-Quran pages to pages without illumination, and to segment Al-Quran pages to text line images without any changes on the content. First we apply a pre-processing which includes binarization. Then, we detect the illumination of Al-Quran pages. In this stage, we introduce the vertical and horizontal white percentages which have been proved efficient for detecting the illumination. Finally, the new images are segmented to text line. The experimental results on several Al-Quran pages from different Al-Quran style demonstrate the effectiveness of the proposed technique.

ABSTRAK

Proses penemberengan Al-Quran memerlukan kajian yang berhati-hati. Ini kerana Al-Quran adalah kitab Allah swt. Sebarang kesalahan penemberengan akan memberikan kesan kepada kesucian Al-Quran. Kesukaran yang dihadapi adalah illuminasi yang mengelilingi kawasan teks Al-Quran dan juga garisan hitam. Pada kajian ini, kami mencadangkan satu algoritma baharu untuk mengenalpasti illuminasi pada setiap muka Al-Quran. Tujuan adalah untuk menembereng Al-Quran muka ke mukadan baris ke baris tanpa mengubah apa-apa pada kandungan Al-Quran. Mulanya, prapemprosesan digunakan dengan menggunakan proses binari. Kemudian, illuminasi dikenalpasti. Pada tahap ini, kami memperkenalkan peratusan menegak dan mendatarberdasarkan kepada piksel putih dapat melaksanakan penemberengan dengan baik. Akhir sekali, imej baru terhasil yang bebas dari illuminasi. Keputusan ujikaji menggunakan stail Al-Quran menunjukkan teknik cadangan adalah efektif.

ACKNOWLEDGEMENT

First and foremost, praise to Allah, for giving me this opportunity, the strength and the patience to complete my project finally, after all the challenges and difficulties. I would like to take this opportunity to express my sincere acknowledgement to my supervisor Dr. Mohd Sanusi Bin Azmi from the Faculty of Information & Communication Technology Universiti Teknikal Malaysia Melaka (UTeM) for her essential supervision, support and encouragement towards the completion of this thesis. Thanks for king Fahd Glorious Quran Printing Complex to publish the styles of Al-Quran to use it during the research.

To my beloved my family and the jewel my heart my mother. Thank you for the sacrifices, patience, support and compassion which has become one enters my life. Not to forget also to all my colleagues and friends struggling for Master's that inspire a vision, guidance and sharing experiences.

Special thanks to all my peers, my father, beloved mother and siblings for their moral support in completing this degree. Lastly, thank you to everyone who had been to the crucial parts of realization of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Image processing is a popular research area in computer science. Today, image processing not only focuses on fundamental issues addressed by researches also the suitability of the research into several domains such as biometric (Phillips et al., 1998), geographical information system (Câmara et al., 1996), character recognition (Omar, 2000), document analysis (Sauvola and Pietikäinen, 2000) and others.

Image Processing is a technique to enhance raw images received from cameras/sensors placed on satellites, space probes and aircrafts or pictures taken in normal day-to-day life for various applications (Rao, 2004). There are three main categories of image processing: Image Enhancement, Image Rectification, and Restoration, and Image Classification.

Various techniques have been developed in Image Processing during the last four to five decades. Most of the techniques are developed for enhancing images obtained from unmanned spacecrafts, space probes and military reconnaissance flights. Image Processing systems are becoming popular due to easy availability of powerful personnel computers, large size memory devices, graphics software etc (Rao, 2004).

Besides, Khairuddin Omar (2010) and Mohammad Faidzul et al. (2010) the image processing body knowledge consists of several phases start from data collection, pre-processing, feature extraction, feature selection, classification, and post processing (Azmi, 2013; Nasrudin et al., 2010; Omar, 2000). Each phase in the image processing has sub-processes. (Omar, 2000) has category the pre-processing phase for the Arabic/Jawi character recognition into Binarization, Edge detection, thinning, and Segmentation before feature extraction took place.

In this research, focus is in segmentation for holy Quran. Thus, Segmentation for holy Quran based on processes for segmenting Arabic/Jawi handwritten texts.

Holy Quran is book of Allah swt. Al-Quran consists of 30 chapters, 114 Surah and 6036 *Ayat*. However, number of pages and lines are difference based on publishers. Table 1.1 below shows example of printed Al-Quran from different publishers.

Table 1.1: Summary of Ayat, Page and Line from printed Al-Quran

Al-Quran (Version)	Ayat	Page	Line per page	Total line
Madinah	6236	604	15	9060
Al-Quran Al-Hakeem	6236	608	15	9120
Al-Quran Al-Kareem	6236	617	15	9225
Al-Quran Al-Majeed	6236	855	13	11115
Mushaf Al-Madinah Quran Majeed {Nastaleeq}	6236	619	15	9285
Mushaf Al-Madinah Quran Majeed	6236	625	15	9375

Based on Table 1.1, issues on segmentation Al-Quran is not uniform. Although the same number of chapter, surah and ayat, but number of page and line are difference. The difference number of page and line become interesting topic to be studied especially in the segmentation process.

Segmentation process for segmenting Al-Quran needs to be studied carefully. This is because Al-Quran is the book of Allah *swt*. Any incorrect segmentation will affect the holiness of Al-Quran.

Currently, exist some segmentation techniques such as Naive Bayes Classifier (Bidgoli and Boraghi, 2010), however the segmentation techniques focus on segmenting object such as text and face segmentation (Khattab et al., 2014). Also, there are some segmentation techniques for Arabic and Jawi character (Omar, 2000). Although, Arabic and Jawi characters are quite near to Al-Quran, however the techniques in Arabic and Jawi do not have diacritical marks (Omar, 2000).

In this research, one technique for segmenting Al-Quran will be proposed. The technique will consider diacritical marks (Tashkil) in order to protect the holiness of Al-Quran.

The propose technique will be evaluated based on comparison with original Al-Quran text. Any missing diacritical marks (Tashkil), words, and sentences will be considered as incorrect evaluation.

1.2 Research Background

Al-Quran is the last book of Allah swt. In Table 1.1, there are difference page and line for each printed Al-Quran. Besides, Figure 1.1, Al-Quran is written with different style of writing and different Illumination. Illumination here is referring to decoration in every page of Al-Quran.

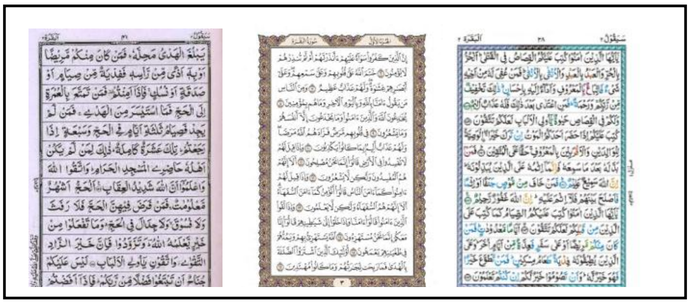

Figure 1.1: Al-Quran writing styles

The segmentation process in this research is use to prepare image for feature extraction process. In the Al-Quran, texts and Diacritical marks (Tashkil) will be extracted. Thus, the illumination and some empty space will be removed.

In this research “illumination” refers to the art of embellishing and decorating the Holy Quran. It is used in Islamic arts. This is used in the first and last page of Al-Quran, a head of each ‘*Sura*’ and a border of each page in Al-Quran and other places (Tajabadi et al., 2009).

Forming decorative arrays and Islamic designs in this holy Quranic art has come from consolidated ideas and worldviews of artists of this field. So that, many calligraphists and number of (not less) illuminators read Quran in a memorized manner. But those less people who were not able to read Quran in a memorized manner, were so familiar with its verses that it had become an integrated part of their nature. (Lings, 1998). Manifestation of spirituality in illumination of Quran is to the extent that it has made this art worthy of companion of the holy Quran and in fact, manifestation of the divine realm is duty of the art evoked by the word of God. And it can be said that Quran itself has opportunities that stimulates religion (Lings, 1998).

1.3 Problem Statement

Arabic language (Quran Language) has markings called “diacritical marks” or “diacritics” represent short vowels or other sounds if one of these diacritical marks ignored it will change the meaning of the word, There are many Arabic character recognition techniques which can recognize the characters of text or the whole page Khairuddin Omar(2000), Mohamad Faidzul (2010) but all these techniques provide recognition for characters without considering the diacritical marks, which may affect the meaning of the Quran's word and the holiness of Al-Quran.

1.4 Research Questions

- How the segmentation process for Al-Quran happen in image processing domain?
- How to segment illumination occur in Al-Quran with different form of illumination and without missing any diacritic?

1.5 Research Objectives

This study has certain objectives as follows:

- To propose framework for Segmenting Al-Quran into page and line
- To propose a technique to segmenting texts of Al-Quran

1.6 Project Significant

The Holy Quran is very important to the Muslims with respect to its authenticity. In this project the removal Illumination technique from the Holy Quran page will be proposed, which enable us remove the Illumination of page according to the percentage of the binary numbers in the Quran page image. The result can be used by the researchers to compare between the copies of the Holy Quran for identifying the originality of copies.

1.7 The Scope of Research

The scope of research is:

i. This research study will primarily focus on removing the border of the Quran pages by cropping the page of Quran first, then crop line by line of the page without the empty space

ii. This technique will be applied on all the Quran pages except the first and second pages which are "Surat Al-Fatihah and the first page of Surat Al-Baqarah" in the Quran that its writing styles comes with a circle border for these two pages.

1.8 Expected Outcomes

The main expected outcomes for this research to design application that provide a technique that produced better image for Holy Quran page image without the border, then cropping the image page line-by-line after line segmentation.

1.9 Conclusion

This thesis addressed the problem of removal image border for sensitive digital Holy Quran pages image. We proposed a very robust and secure approach against the remove border without any changes on the content of the Holy Quran pages to maintain the authenticity of the Quran text-image content. Our objective is to segment the page of the Quran into lines and also to remove Illumination.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Nowadays we are living in a world that is almost entirely digital. So, many digital documents are available. Research on document originality are exist (Qadir and Ahmad, 2006). For this research, study will be done on Al-Quran. There are too many printed versions of Al-Quran as well as digital copy. This research will do the segmentation on Al-Quran for the removing illumination and segmenting Al-Quran by lines. This research will prepare the Al-Quran for feature extraction phase and the final aim is to validate the originality of Al-Quran.

There are too many research on segmentation, however, segmenting the Al-Quran need to do carefully in order to preserve the holiness of Al-Quran. Research that nearly to Al-Quran are Arabic/Jawi segmentation. But, both are not suitable to be applied for Al-Quran due to diacritics diacritical marks (*Tashkil*), words, and sentences.

The first step before segmentation phase is preprocessing documents to produce a clean image of the document. Al-Quran page image contained Illumination. Detecting and removing these unwanted areas is critical to achieve better text segmentation results. Before the illumination detection and removal takes place, we first proceed to image binarization using the efficient technique proposed in (Gatos et al., 2006).

There are some segmentation techniques for Arabic and Jawi character (Omar, 2000). (Omar, 2000) has category the pre-processing phase for the Arabic/Jawi character recognition into Binarization, Edge detection, thinning, and Segmentation before feature extraction took place.

We are trying in our research to understand the image processing and text segmentation. Images used in our study will be are an image of Al-Quran pages. At this point, Will be explained the previous studies on text line segmentation in detail. The main objective of this research is to seek out the better method to segment Al-Quran pages without missing any character or diacritical marks.

2.2 Image Pre-processing

2.2.1 Progress in Binarization Studies

A binary image (Stathis et al., 2008) is a digital image that has just two feasible values meant for every pixel. Normally, two colors are used for a binary image i.e. black and white however any two colors can be used. The color used for the objects in the image is the foreground color while the rest of the image is the background color. Binary images (Su et al., 2011) frequently occur in image processing as masks or as the outcome of some operations as segmentation and thresholding. Few input/output devices, for example, laser printers, bi-level computer displays, are able to just handle bi-level images. Binary images are formed from color images by segmentation.

Various approaches as well as techniques were developed to improve documents images quality. Binarization is one of the most important pre-processing steps which consist to separate foreground and background of documents images. It converts a gray-scale document image into a binary document image.

Image binarization is typically executed in the preprocessing phase of image processing of several documents. It is the process of separation of pixel values into dual collections, black as foreground and white as background.

Mohd Sanusi (2013) shows the processes within the pre-processing carried out by researchers within the Jawi script. This process is predicated on the process performed by Khairuddin Omar (2000). Overall stages that were utilized by the pioneers of Jawi script Khairuddin Omar (2000) has committed method of images reborn to a scale illustration of color, skew and slant correction, noise removal, and thinning of the frame (skeleton). However, the process utilized by Khairuddin Omar (2000) not all of them are utilized by researchers to conduct pre-processing as shown in Table 2.1.

Table 2.1: Image processing steps for Jawi pattern recognition (Azmi, 2013)

Researcher for previous studies	Format Conversion	Skew and slant correction	Noise Removal	Thinning
Khairuddin Omar (2000)	√	√	√	
Mazani Manaf (2002)	√		√	√
Mohammad Roslim (2002)				√
Mohammad Faidzul (2010)	√		√	

By Refer to Table 2.1, Khairuddin Omar (2000) during the pre-processing phase has been accomplished on the scale of the transformation of binary format. In addition to that, Mazani Manaf (2002) and Mohammad Faidzul (2010) come out to gray scale format during the transformation. After that, by these researchers has carried out the methodology of noise removal. Khairuddin Omar (2000) in the noise removal process they use Median Filter and then he has been using the technique proposed by Sharaf El-Deen et al. (2003) to "change the image format to a binary scale". After this, using gradient orientation histogram implements the skew as well as slant correction image. The ultimate procedure carried out through Khairuddin Omar (2000) the thinning skeleton did in the pre-processing. The algorithms used by Khairuddin Omar (2000) for the thinning skeleton using sequential thinning algorithm Safe-Point Thinning Algorithm (SPTA) proposed by Naccache, & Shinghal (1984). SPTA algorithm is also used by Roslim Mohammad (2002) in the thinning of the Jawi script.

Mazani Manaf (2002) in pre-processing employs "gamma correction and intensity". Next the uses of a linear function with the planned technique desires by Parker (1994) to convert the image to grayscale format. The noise removal method, then being executed by using erosion operation followed by reclamation as proposed by Zhang, & Suen (1984). After that, to de-noise employs the median filter. Finally, Mazani Manaf (2002) using a simple sequential thinning algorithm performs a thinning process framework based on Zhang and Suen (1984).

Mohammad Faidzul (2010) in his study has taken knowledge from nine writers. The primary method done by him is doing segmentation that performed manually. however he did't doing the noise removal in his research. The results of this manually segmentation, he obtained 993 total characters and 540 sub-word from nine last author. Next, format conversion method to be enforced using the binary scale thresholds of 127. Mohammad Faidzul (2010) did't perform the repair method skew and slant and additionally the thinning skeleton.

Abdenour Sehad et al.(2013) (Sehad et al., 2013) has present a capable scheme for binarization of ancient and degraded document images, grounded on texture qualities. The suggested technique is an adaptive threshold-based. It has been calculated by using a descriptor centered on a co-occurrence matrix and the scheme is verified objectively, on DIBCO dataset degraded documents furthermore subjectively, utilizing a set of ancient degraded documents offered by a national library. The outcomes are acceptable and assuring, present an improvement to classical approaches.

Hossein Ziaei Nafchi et al.(2013) (Nafchi et al., 2013) has concluded that the pre-processing and post processing phases meaningfully advance the performance of binarization approaches, particularly in the situation of harshly degraded ancient documents. An unverified post processing technique is presented founded on the phase-preserved denoised image and also phase congruency features extracted from the input image. The central part of the technique comprises of two robust mask images that can be used to cross the false positive pixels on the production of the binarization technique. Firstly, a mask with an extreme recall value is attained from the denoised image with the help of morphological procedures. In parallel, a second cover is acquired dependent upon stage congruency features. At that point, a median filter is utilized to evacuate noise on these two masks, which then are utilized to rectify the yield of any binarization strategy.

Jon Parker et al.(2013) (Parker et al., 2013) has studied that regularly documents of notable noteworthiness are ran across in a state of deterioration. Such archives are regularly examined to all the while history and announce a disclosure. Changing over the data found inside such reports to open information happens all the more rapidly and inexpensively if a programmed technique to upgrade these corrupted archives is utilized as opposed to improving each one document image by hand. A novel mechanized image upgrade approach that indulges no preparation information was introduced. The methodology was valid to images of typewritten text in addition to hand written text or both.

Konstantinos Ntirogiannis et al.(2013) (Ntirogiannis et al., 2013) has analysed that document image binarization is of incredible value in recognition pipeline and document image examination as it disturbs further phases of the recognition procedure. The assessment of a binarization technique helps in examining its algorithmic conduct, and also confirming its adequacy, by giving qualitative and quantitative sign of its execution. A