QURANIC DIACRITIC AND CHARACTER SEGMENTATION AND RECOGNITION USING FLOOD FILL AND K-NEAREST NEIGHBORS ALGORITHM

FAIZ E A L ALOTAIBI

FSKTM 2019 59

QURANIC DIACRITIC AND CHARACTER SEGMENTATION AND RECOGNITION USING FLOOD FILL AND K-NEAREST NEIGHBORS ALGORITHM

By

FAIZ E A L ALOTAIBI

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

September 2019

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

QURANIC DIACRITIC AND CHARACTER SEGMENTATION AND RECOGNITION USING FLOOD FILL AND K-NEAREST NEIGHBORS ALGORITHM

By

FAIZ E A L ALOTAIBI

September 2019

Chairman : Associate Professor Muhammad Taufik Abdullah, PhD
Faculty : Computer Science and Information Technology

The detection, recognition and conversion of the characters in an image into a text are called optical character recognition (OCR). A distinctive type of OCR is used to process Arabic characters, namely, Arabic Optical Character Recognition (AOCR). OCR is increasingly used in many applications, where this process is preferred to automatically perform a process without human intervention.

The Quranic handwriting text contains two elements, namely, diacritics and characters. However, the current Arabic handwritten OCR system produces low levels of accuracy and no research focused on Quran image recognition.

The current AOCR inaccurately recognizes diacritic and characters, and the research and efforts in the area of AOCR are insufficient. Many studies have been carried out so far, but for Quran handwriting has not been researched as thoroughly as Arabic, Latin or Chinese handwritten systems. The current research is focused on solving the mentioned problems through improving the accuracy of recognition rate of AOCR by proposing a new segmentation, feature extraction methods and finding a suitable classification.

In this thesis, a new techniques, methods and algorithms are proposed to check the similarities and originalities of the Quranic handwriting content. The diacritic detections are performed using a region-based algorithm with 89% accuracy and 95% improved by using flood fill segmentations method. 2DMED feature extraction accuracy was 90% for diacritics and 96% improved by applied CNN. Character recognition is performed based on the projection method with 86% accuracy, and 92%

improved by using flood fill. 2DMED in characters was 88% and 91 % after improved by applied CNN. For classification, KNN used before and after enhancement technique based on essential vector with our dataset, the diacritic accuracy was 96.4286% after enhancement, which is better than the 87.5020% in detecting. For characters was at 92.3077% improvement, which is better that normal KNN algorithm which exhibited an 86.1429% accuracy in detecting.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENSEGMENAN DAN OENGECAMAN DIAKRITIK DAN AKSARA AL-QURAN MENGGUNAKAN ISI BANJIR DAN K-NEAREST NEIGHBORS ALGORITHM

Oleh

FAIZ E A L ALOTAIBI

September 2019

Pengerusi : Profesor Madya Muhammad Taufik Abdullah, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Pengesanan, pengecaman dan penukaran aksara dalam imej kepada teks dipanggil pengecaman aksara optik (OCR). Jenis OCR yang tersendiri digunakan untuk memproses aksara Arab, iaitu pengecaman aksara Optik Arab (AOCR). OCR semakin banyak digunakan dalam banyak aplikasi, di mana proses ini lebih disukai untuk melakukan proses secara automatik tanpa intervensi manusia.

Teks tulisan tangan Al-Quran mengandungi dua unsur, iaitu, diakritik dan aksara. Walau bagaimanapun, system ocr tulisan tangan bahasa Arab semasa menghasilkan ketepatan yang rendah dan tiada penyelidikan yang memberi tumpuan kepada pengecaman imej Al-Quran.

AOCR semasa tidak dapat mengecam diakritik dan aksara dengau tepat, dan penyelidikan dan usaha dalam bidang AOCR tidak mencukupi. Banyak kajian telah dilakukan setakat ini, tetapi bagi Al-Quran tulisan tangan belum dikaji secara menyeluruh seperti sistem tulisan tangan Arab, Latin atau Cina. Penyelidikan semasa memberi tumpuan kepada menyelesaikan masalah yang dingatakan melalui peningkatan ketepatan kadar pengecaman AOCR dengan mencadangkan satu segmentasi baru, kaedah pengekstrakan ciri dan mencari klasifikasi yang sesuai.

Dalam tesis ini, teknik, kaedah dan algoritma baru dicadangkan untuk memeriksa persamaan dan keaslian kandungan tulisan tangan Al-Quran. Pengesanan diakritik dilakukan menggunakan algoritma berasaskan rantau dengan ketepatan 89% dan 95% ditingkatkan dengan menggunakan kaedah segmentasi isi banjir. Ketepatan pengekstrakan ciri 2DMED adalah 90% untuk diacritics dan 96% ditingkatkan dengan

menggunakan CNN. pengecaman aksara dilakukan berdasarkan kaedah unjuran dengan ketepatan 86%, dan 92% ditingkatkan dengan menggunakan isi banjir. 2DMED untuk aksara adalah 88% dan 91% selepas diperbaiki dengan CNN. Mengguuakan Untuk klasifikasi, KNN menggunakan sebelum dan selepas teknik peningkatan berdasarkan vektor penting dengan dataset kami, ketepatan diakritik adalah 96.4286% selepas peningkatan, yang lebih baik daripada pengesanan 87.5020%. Untuk aksara dengan peningkatan 92.3077%, yang lebih baik daripada algoritma KNN biasa yang memperlihatkan ketepatan 86.1429% dalam mengesan.

ACKNOWLEDGEMENTS

In the Name of Allah The Most Benevolent, The Most Merciful

I am grateful to Allah, for allowing me to complete my work and produce this thesis. The completion of this thesis would not be possible without the help and support of many people.

Firstly, I would like to thank my parents for their full support and patience throughout the process of carrying out the research leading this thesis. I am also grateful to have brothers and sisters who always understood my obligations in spending time doing the research, taking away much of my attention and time for them. I thank them for all their love and continuous encouragement.

Secondly, I would also like to thank my wife for her love and constant support. Without you I would not be the person I am today. I owe you everything.

Lastly, I would also like to attribute special thanks to my supervisor, Assoc. Prof. Dr. Muhamad Taufik Abdullah, for allowing me to strive and prosper under his guidance and supervision for many years. Similar thanks to all my co-supervisors - Prof. Dr Rusli Haji Abdullah and Prof. Dr Rahmita Wirza. Also, many thanks go to all my seniors and colleagues in the Faculty of Computer Science and Information Technology (FSKTM), Universiti Putra Malaysia (UPM), who have regularly lent a helping hand.

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software

Signature: _____ Date: _____

Name and Matric No. : Faiz E A L Alotaibi, GS43095

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) were adhered to.

Signature: _____
Name of Chairman of Supervisory Committee: Associate Professor Dr. Muhamad Taufik Abdullah

Signature: _____
Name of Member of Supervisory Committee: Professor Dr. Rusli Haji Abdullah

Signature: _____
Name of Member of Supervisory Committee: Professor Dr. Rahmita Wirza O.K. Rahmat

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2DMED	Two-Dimensional Maximum Embedding Difference
ACSA	Arabic Character Segmentation Algorithm
ANN	Artificial Neural Network
AOCR	Arabic Optical Character Recognition
CC	Connected Component
CNN	Convolutional Neural Network
COG	Center of Gravity
EV	Essential Vector
ERM	Empirical Risk Minimization
FD	Fisher's Discriminant
FF	Flood Fill
FN	False Negative
FP	False Positive
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
KNN	K-Nearest Neighbors
LGH	Local Gradient Histogram
OCR	Optical Character Recognition
OPF	Optimum-Path Forest
PCA	Principal Component Analysis
PQ	Power Quality
RGB	Red Green Blue Color Model
SOM	Self-Organizing Map
SSDA	Stacked Sparse Denoising Auto-encoder
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VCD	Vapnik–Chervonenkis Dimension

CHAPTER 1

INTRODUCTION

1.1 Introduction

An Optical Character Recognition (OCR) is the process of converting an image representation of a document into an editable format (Al-Shatnawi, 2012). OCR applications enable users to search for documents stored in the format of images by converting them into text, which are easily processed by computers.

Each OCR system contains a few processing stages, where a particular task is accomplished in each stage and that the output of each stage is considered as the input for the next stage. Typically, an OCR system consists of a few main stages which includes preprocessing, segmentation, feature extraction, and classification. However, despite decades of intensive investigation and research, the ultimate goal of developing a method or an OCR system that has the same reading capabilities as humans has yet to be achieved.

This is particularly true for Arabic Optical Character Recognition (AOCR). The Al-Muhtaseb et al. (2008) provided a survey on AOCR with respect to text recognition by focusing on the characteristics and technologies of text recognition in the Arabic language. The result of the survey instigates a motivation for further query into this area. In addition, the abundance of applications generating texts and images - an important trend in the current literature – renders this effort pertinent.

In this thesis, a method to identify Quranic diacritics and characters segmentation is presented. The research on handwritten Quranic text recognition is challenging yet it gains ever more attention due to the increasing usage of hand-held devices such as computers, digital notebooks, and advanced cellular phones.

Some techniques have been used to build several handwriting recognition systems, such as Neural Networks, Hidden Markov Model and Fuzzy Logic (Schiuma, 2012). After 1990, complex character recognition algorithms were developed. Many recognizers used these same sophisticated methodologies, not to mention natural language processing techniques. However these current algorithms have high error rates, inhibiting the user from achieving full accurate recognition, and in addition to that, comparisons with the Quran. The current AOCRs are not accurate in recognizing the diacritic and characters in the Quran, and that stems from the lack of research and effort done in the AOCR area.

This research ventures into that area and need, focusing on Quranic characters and diacritics recognition.

1.2 Quranic Characters Recognition

The Arabic language is considered a dominant language in the Arabic diaspora. It is the second most spoken language in a population of approximately 280 million people all over the world (Versteegh, 2014). Based on a recent study (Hakak, 2017), the Arabic language is ranked fifth as the most common language used in the world.

Religious beliefs and practices of Islam demand Muslims to read the holy Quran using the Arabic language and to have basic knowledge of the Arabic language so that they may understand what they recite during their prayers. Moreover, there are many other languages associated with the Arabic language which have similar characters, such as Persian, Jawi, Pashto, Urdu, Bengali, etc. (Abuzaraida, 2010). Some images of the Quran found online contain diacritics that are difficult to identify - inadvertently altered diacritics due to stylistics or by how different screens display the information differently, or how the website was coded, as shown in Figure 1.1.

Figure 1.1 : Less diacritics (left), many diacritics (right)

People have the ability to recognize characters without major difficulty, reading papers or books with different prints, sizes and orientations. However, developing an OCR system that has the same ability to read and recognize Arabic Quranic characters as a human has remained far from reach (Shanthi, 2017).

There are two types of OCR systems found in literature - Typewritten and Handwritten (Hamdi, 2014). The Typewritten OCR is mainly purposed to identify documents which are typed and scanned. Handwritten OCR is used to recognize text that are written by human hands. The main difference between these two systems is that the Typewritten OCR is simpler in terms of design. Furthermore, the recognition rate of the Typewritten OCR is higher. A recent study conducted by Razak (2008) provided a comprehensive overview of the method of off-line handwriting text line segmentation. The authors highlighted that the characteristics of text line structure in handwritten documents and also text segmentation are the pertinent research challenges.

1.3 Quranic Diacritics Recognition

The Arabic alphabets is a widely used alphabetic writing system in the world (Versteegh, 2014), containing 28 basic characters. The alphabets was first used to write texts in Arabic, most notably the Quran (the holy book of Islam). With the spread of Islam, it came into use to write many languages at various times, such as Urdu, Pashto, Uyghur (in China), not to mention Ottoman Turkish and Spanish in Western Europe (Beeston, 2016).

To accommodate the needs of these languages, new letters and symbols were added to the original alphabets. This process is known as the Ajami transcription system, which is different from the original Arabic alphabet. Then many modifications and improvements have been made to the Arabic writing script, which results in additional letters and strokes. The new strokes are called diacritics, and the purpose of adding these diacritics was:

1. To distinguish between letters of the same or similar shape.
2. To indicate sounds (vowels and tones) that are not conveyed by the basic alphabets

1.4 Statement of Problem

Since the Arabic language is a significant language - ranking number Fifth as one of the most widely used languages in the world – the need to its proficiency is paramount. Religious beliefs and practices of Islam demand Muslims to be able to read the holy Quran using the Arabic language. Muslims also require a basic knowledge of Arabic during their prayers (Fan et al., 2017). There are also many other languages associated with the Arabic language which share similar characters such as Persian, Jawi, Pashto, Urdu, Bengali, etc. (Zeki, 2010).

Errors frequently occur from Arabic letters associated with diacritics. It believes that the separation of the identification of diacritics from letters will highly improve the accuracy of the results when processing any Arabic text image.

Moreover, there are lots of minor problems that occur in every identification process such as the extraction of overlapping Arabic characters in an image and text retrieval from images using different writing styles. This area needs further research in order to identify the issues that cause misreads and incorrect identification. Possible solutions may lie in the use of segmentation techniques and machine learning approaches (Hakak, 2017).

With the massive growth of mobile applications and websites that contain Quranic verses made available, it is a challenge to identify the authenticity of a digital copy of the Quran, or even it's verses. Due to the sensitivity and the nature of the Quran, even a small change is intolerable as it could result in a completely different reading and change its meaning. As the stated problem size continues to scale up with the exponential development of the Internet and its content, it is essentially paramount to research on the classification of Quran authenticity. Existing Quran classification is mostly focused on different aspects of the Quran, such as Quranic themes, subjects or topics. There is a clear lack of research done that has focused on the classification of Quran images (Ridzuan, 2017).

Many experiments conducted were based on different subsets of features. Typically, a web-based prototype is developed. Future work should focus on applying more machine learning and optimization techniques in order to achieve higher evaluation measurements and incorporate these methods to improve the detection and authentication of Quranic verses from images (Sabbah & Selamat, 2014).

The K-Nearest Neighbor is a non-parametric type of algorithm, meaning it doesn't have to create an assumption about its environment. The number of parameters depends on the number of training data. Classification is done by calculating the average/majority distance from a test vector to its neighboring training vectors. KNN can better classify letters which look similar to each other, capital and small letter at image (Ong & Suhartono, 2016).

Research by Adeleke et al. (2017), applied three text classification algorithms namely, k-Nearest Neighbors, Support Vector Machine, and Naïve Bayes. The problem for this research on labelling the dataset, the result shows 70% accuracy.

On top of the two common OCR systems found in literature, namely the handwritten and typewritten, OCRs can be further classified by two other categories – the Online recognition and the offline recognition systems (Abdi & Khemakhem, 2015).

The image of the typewritten or handwritten text is created through scanning using offline. However, for the online, it uses devices like a phone or a portable personal computer to create an image that is the input of the OCR system. Then, OCR system reads the image and is analyzed for recognition. Research has been conducted on various languages such as Japanese, Chinese, and Latin characters, in which most of them function based on the premise that they isolate characters individually when crafting OCR algorithms. Nevertheless, this may not be appropriate for the languages with cursive scripts such as Arabic.

All the above presents a complex challenge that is the Arabic optical character recognition. Consequently, few studies have been done on the Arabic OCR and its character recognition compared to languages like Latin, Chinese and others (Mesleh et al, 2012). Problem can summarize based on above researcher on these steps:

- Quran handwritten has cursive overlapping
- Shape of letters can be solved on preprocessing and segmentation stages.
- OCR classification still so poor.

1.5 Research Objectives

The main objective of this study is to improve recognition rate accuracy of Quran handwritten word and diacritics recognition. To achieve the objective, this research proposes three following methods:

1. To segmentation Quran diacritics and characters handwriting using flood fill.
2. To extract feature extraction method for Quranic character and diacritics recognition classified by CNN.
3. To compare with normal KNN and after applying new KNN enhancement technique on our dataset.
4. To improve Arabic framework for Quran diacritics and characters handwriting.

1.6 Scope of Research

The Othman-Taha is one of the most used fonts for the holy Quran. Mushaf al-Madinah is used as the standard version of the Quran. The proposed method will develop a similarity check algorithm for Quranic characters and diacritics recognition in images based on these fonts and versions.

In this research, a new method to check the similarity and originality of Quranic content is proposed. This method consists of Quranic diacritics and characters recognition techniques. The diacritics detection will be performed using the region based algorithm, which will divide the text to rows, before identifying the baseline row by using other two rows to find the upper and lower baselines. This will make it much easier to locate a pixel (Saeed & Albakoor, 2009). An optimization methods are applied to increase the recognition ratio.

The character recognition method is performed based on the projection method. An optimization method are applied to increase the recognition ratio. Then the combination of the result of these two methods will enabled us to make the comparison with the standard Mushaf al Madinah with our dataset as the benchmark and find matches on the Quran.

Lastly, the similarity ratio of the given image and its matching benchmark is determined.

1.7 Research Contribution

The new method proposed in this research to identify the similarity and originality of Quranic content using improved KNN algorithm. This thesis makes four contributions to research knowledge in the Quran handwritten diacritic and character recognition techniques

1. Enhance the accuracy of diacritics and characters by apply new segmentations methods. These methods include flood fill and clustering segmentations.
2. Enhance the accuracy of recognition rate of Quran handwritten by apply new feature extraction method. These methods include diacritics, capital word and small word on characters.
3. Enhance the accuracy rate by compare this experiment with normal KNN and KNN enhancement technique.
4. Create dataset for Quran handwritten diacritics and characters.
5. Enhance Quran handwriting diacritics and characters framework.

1.8 Organization of the Thesis

Chapter 2 aims to survey the research undertaken in the field of diacritics and characters recognition. The chapter provides knowledge of image processing and identifies key issues with respect to Quran handwritten diacritics and characters recognition. This chapter highlights the machine learning algorithms for image recognition and investigates the related diacritics and characters algorithms. It will also discuss several approaches to the recognition problem. These approaches consider different scenarios, which consider the application types, the datasets, the type of algorithms used and the various constraints that might be imposed. Moreover, the chapter discusses optimization by means of placing focus on machine algorithms like KNN, ANN and SVM, etc.

Chapter 3 presents a review of the research conducted within the image recognition in solving problems of Quran handwrittenic diacritics and characters. It discusses"proposed solution based Quran handwrittenic Diacritics Recognition in image processing, Quran handwrittenic Character Recognition in Image, and Checking Similarity Matching with Standard Version of Quran handwritten."

Chapter 4 presents a new proposed method which tries to identify the importance of proposed model development and its algorithms.

Chapter 5 presents a proposed method which places importance in the design of a systematic evaluation procedure in order to provide a verification of its use. This chapter offers performance evaluation and statistical modeling based on the proposed algorithm which aims to compare it to the performance of other algorithms.

First, the chapter provides a description of benchmarks that were used for the evaluation of the proposed algorithm. Second, the simulation environment is described in detail. Then, statistical analysis that is derived for validation of the findings is described. A series of experiments to show platform-independence of our proposed solution is described and the comparative study that is designed to demonstrate the proposed algorithm is described.

Chapter 6 concludes the major contributions of the thesis. It also outlines the limitations and opportunities to further improve or extend the work presented in the thesis. To this end, this thesis stands as a substantial effort to optimize image recognition.

REFERENCES

Abdi, M. N., & Khemakhem, M. (2015). A model-based approach to offline text-independent Arabic writer identification and verification. *Pattern Recognition*, 48(5), 1890-1903.

Abudena, M. A., & Hameed, S. A. (2015). Toward a novel module for computerizing Quran's full-script writing. *International Journal of Computer Systems*, 2(11), 469-474.

Aburas, A. A., & Rehiel, S. M. (2007). OffLine Omni-Style Handwriting Arabic Character Recognition System Based on Wavelet Compression. *Arab Research Institute in Sciences & Engineering*, 3(4), 123-135.

Abuzaraida, M. A., Zeki, A. M., & Zeki, A. M. (2010, December). Segmentation techniques for online Arabic handwriting recognition: a survey. *In Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World (ICT4M) 2010* (pp. D37-D40).

Adeleke, A. O., Samsudin, N. A., Mustapha, A., & Nawi, N. M. (2017). Comparative analysis of text classification algorithms for automated labelling of Quranic verses. *Int. J. Advanc. Sci. Eng. Info. Tech*, 7, 1419-1427.

Aizenberg, I., Aizenberg, N., Hiltner, J., Moraga, C., & Zu Bexten, E. M. (2001). Cellular neural networks and computational intelligence in medical image processing. *Image and Vision Computing*, 19(4), 177-183.

Akiyama, T., & Hagita, N. (1990). Automated entry system for printed documents. *Pattern recognition*, 23(11), 1141-1154.

Aksoy, S. (2010). Introduction to pattern recognition. *Lecture Notes Comput. Sci*., CS, 551.

Al-Badr, B., & Haralick, R. M. (1998). A segmentation-free approach to text recognition with application to Arabic text. *International Journal on Document Analysis and Recognition*, 1(3), 147-166.

Al-Dmour, A., & Fraij, F. (2014). Segmenting Arabic handwritten documents into text lines and words. *International journal of advancements in Computing technology*, 6(3), 109.

Ali, A., Ahmad, M., Rafiq, N., Akber, J., Ahmad, U., & Akmal, S. (2004, December). Language independent optical character recognition for handwritten text. *In 8th International Multitopic Conference, 2004. Proceedings of INMIC 2004*. (pp. 79-84).

AlKhateeb, J. H., Khelifi, F., Jiang, J., & Ipson, S. S. (2009, November). A new approach for off-line handwritten Arabic word recognition using KNN classifier. *In 2009 IEEE International Conference on Signal and Image Processing Applications* (pp. 191-194).

Al-Ma'adeed, S. A. (2004). *Recognition of off-line handwritten Arabic words* (Doctoral dissertation, University of Nottingham).

Almeida, L. G., Backović, M., Cliche, M., Lee, S. J., & Perelstein, M. (2015). Playing tag with ANN: boosted top identification with pattern recognition. *Journal of High Energy Physics*, 2015(7), 86.

Al-Muhtaseb, H. A., Mahmoud, S. A., & Qahwaji, R. S. (2008). Recognition of off-line printed Arabic text using Hidden Markov Models. *Signal processing*, 88(12), 2902-2912.

Al-Shatnawi, A. M. (2012). A new method in image steganography with improved image quality. *Applied Mathematical Sciences*, 6(79), 3907-3915.

Álvarez-Meza, A., Valencia-Aguirre, J., Daza-Santacoloma, G., & Castellanos-Domínguez, G. (2011). Global and local choice of the number of nearest neighbors in locally linear embedding. *Pattern Recognition Letters*, 32(16), 2171-2177.

Al-Yousefi, H., & Udpa, S. (1992). Recognition of Arabic characters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8), 853-857.

Artigas-Fuentes, F. J., Gil-García, R., Badía-Contelles, J. M., & Pons-Porrata, A. (2010). Fast k-NN classifier for documents based on a graph structure. *In Iberoamerican Congress on Pattern Recognition* (pp. 228-235). Springer, Berlin, Heidelberg.

Astudillo, C. A., & Oommen, B. J. (2013). On achieving semi-supervised pattern recognition by utilizing tree-based SOMs. *Pattern Recognition*, 46(1), 293-304.

Atallah, A. S., & Omar, K. (2008). Methods of Arabic language baseline detection–the state of art. *IJCSNS*, 8(10), 137.

Atena F., Abdolhossein S., and Jamshid S. (2013). "Document Image Noises and Removal Methods", *International MultiConference of Engineers and Computer Scientists*, pp. 1-5, 2013

Ayyalasomayajula, K. R., Nettelblad, C., & Brun, A. (2016). Feature evaluation for handwritten character recognition with regressive and generative Hidden Markov Models. *In International Symposium on Visual Computing* (pp. 278-287). Springer, Cham.

B. Bataineh, S. Abdullah, K. Omar, (2011) A statistical global feature extraction method for optical font recognition, in: *Intelligent Information and Database Systems*, vol. 6591 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg, 2011, pp. 257–267.

Beeston, A. F. L. (2016). *The Arabic language today*: Routledge.

Bennet, J., Ganaprakasam, C., & Kumar, N. (2015). A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. *International Arab Journal of Information Technology* (IAJIT), 12.

Bernard, J., Chang, T.-W., Popescu, E., & Graf, S. (2017). Learning style Identifier: Improving the precision of learning style identification through computational intelligence algorithms. *Expert Systems with Applications*, 75, 94-108.

Bokser, M. (1992). Omni document technologies. *Proceedings of the IEEE*, 80(7), 1066-1078.

Bolivar-Cime, A., & Marron, J. S. (2013). Comparison of binary discrimination methods for high dimension low sample size data. *Journal of Multivariate Analysis*, 115, 108-121.

Bonissone, P., Cadenas, J. M., Garrido, M. C., & Díaz-Valladares, R. A. (2010). A fuzzy random forest. *International Journal of Approximate Reasoning*, 51(7), 729-747.

Bouchiareb, F., Bedda, M., & Ouchetai, S. (2006). New preprocessing methods for handwritten Arabic word. *Asian Journal of Information Technology*, 5(6), 609-613.

Boukharouba, A. (2017). A new algorithm for skew correction and baseline detection based on the randomized Hough Transform. *Journal of King Saud University-Computer and Information Sciences*, 29(1), 29-38.

Cervantes, J., Li, X., Yu, W., & Li, K. (2008). Support vector machine classification for large data sets via minimum enclosing ball clustering. *Neurocomputing*, 71(4), 611-619.

Chaudhuri, K., Kakade, S. M., Netrapalli, P., & Sanghavi, S. (2015). Convergence rates of active learning for maximum likelihood estimation. *In Advances in Neural Information Processing Systems* (pp. 1090-1098).

Cherkassky, V., Friedman, J. H., & Wechsler, H. (Eds.). (2012). *From statistics to neural networks: theory and pattern recognition applications* (Vol. 136). Springer Science & Business Media.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 493-507.

Cheung, A., Bennamoun, M., & Bergmann, N. W. (2001). An Arabic optical character recognition system using recognition-based segmentation. *Pattern recognition*, 34(2), 215-233.

Chen, Y., Jiang, H., Li, C., Jia, X., & Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10), 6232-6251.

Chherawala, Y., Roy, P. P., & Cheriet, M. (2013). Feature design for offline Arabic handwriting recognition: handcrafted vs automated. *In 2013 12th International Conference on Document Analysis and Recognition* (pp. 290-294). IEEE.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *In Advances in neural information processing systems* (pp. 577-585).

Cosma, G., Brown, D., Archer, M., Khan, M., & Pockley, A. G. (2017). A survey on computational intelligence approaches for predictive modeling in prostate cancer. *Expert systems with applications*, 70, 1-19.

Dai, L., Hu, H., Chen, Y., & Zhou, M. (2016). Millimeter-wave image target recognition based on the combination of shape features. *In 2016 IEEE International Conference on Information and Automation (ICIA)* (pp. 1732-1736). IEEE.

Daza-Santacoloma, G., Acosta-Medina, C. D., & Castellanos-Domínguez, G. (2010). Regularization parameter choice in locally linear embedding. *Neurocomputing*, 73(10), 1595-1605.

Dehghan, M., Faez, K., Ahmadi, M., & Shridhar, M. (2001). Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM. *Pattern Recognition*, 34(5), 1057-1065.

Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., & Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. *In Advances in neural information processing systems* (pp. 1269-1277).

Dhall, A., Asthana, A., Goecke, R., & Gedeon, T. (2011). Emotion recognition using PHOG and LPQ features. *In Face and Gesture* (pp. 878-883). IEEE.

Dharani, T., & Aroquiaraj, I. L. (2013). Content Based Image Retrieval System Using Feature Classification with Modified KNN Algorithm. *arXiv preprint arXiv*:1307.4717.

Di Martino, M., Hernández, G., Fiori, M., & Fernández, A. (2013). A new framework for optimal classifier design. *Pattern Recognition*, 46(8), 2249-2255.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

El Kessab, B., Daoui, C., Bouikhalene, B., & Salouan, R. (2015). A Comparison between the Performances of Several Distances for Isolated Handwritten Arabic Numerals Recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(6), 9-14.

Fan, Z., Bi, D., He, L., Shiping, M., Gao, S., & Li, C. (2017). Low-level structure feature extraction for image processing via stacked sparse denoising autoencoder. *Neurocomputing*, 243, 12-20.

Fernández, A., Gómez, Á., Lecumberry, F., Pardo, Á., & Ramírez, I. (2015). Pattern recognition in latin America in the "big data" era. *Pattern Recognition*, 48(4), 1185-1196.

Fraiman, R., & Pateiro-López, B. (2012). Quantiles for finite and infinite dimensional data. *Journal of Multivariate Analysis, 108*, 1-14.

Franco-Arcega, A., Carrasco-Ochoa, J. A., Sánchez-Díaz, G., & Martínez-Trinidad, J. F. (2011). Decision tree induction using a fast splitting attribute selection for large datasets. *Expert Systems with Applications*, 38(11), 14290-14300.

Gao, K., Zhu, P., Hu, Q., & Zhang, C. (2016). Unsupervised Subspace Learning via Analysis Dictionary Learning. *In Chinese Conference on Biometric Recognition* (pp. 556-563). Springer, Cham.

Gonzalez, E. C., Figueroa, K., & Navarro, G. (2008). Effective proximity retrieval by ordering permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9), 1647-1658.

Gupta, S., Rana, S., Saha, B., Phung, D., & Venkatesh, S. (2016). A new transfer learning framework with application to model-agnostic multi-task learning. *Knowledge and information systems*, 49(3), 933-973.

Gutub, A. A. A., Al-Alwani, W., & Mahfoodh, A. B. (2010). Improved method of Arabic text steganography using the extension 'Kashida'character. *Bahria University Journal of Information & Communication Technology*, 3(1), 68-72.

Hakak, S., Kamsin, A., Tayan, O., Idris, M. Y. I., Gani, A., & Zerdoumi, S. (2017). Preserving content integrity of digital holy quran: Survey and open challenges. *IEEE Access*, 5, 7305-7325.

Hamdi, H., & Maher, K. (2014, February). Pattern Recognition System based on Distributed Computing Architectures: Clusters, Peer to Peer and Data grid. *In Accepted paper in the Third International Conference on Natural Language Processing (NLP, 2014), Pullman, Sydney, Australia.*

Hein, M., Lugosi, G., & Rosasco, L. (2016). Mathematical and Computational Foundations of Learning Theory (Dagstuhl Seminar 15361). *In Dagstuhl Reports* (Vol. 5, No. 8). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Imani Ma, G., Xu, Z., Zhang, W., & Li, S. (2015). An enriched K-means clustering method for grouping fractures with meliorated initial centers. *Arabian Journal of Geosciences*, 8(4), 1881-1893.

Jayech, K., Mahjoub, M. A., & Amara, N. E. B. (2016). Arabic handwritten word recognition based on dynamic bayesian network. Int. *Arab J. Inf. Technol*., 13(6B), 1024-1031.

Kesidis, A. L., Galiotou, E., Gatos, B., & Pratikakis, I. (2011). A word spotting framework for historical machine-printed documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2), 131-144.

Khan, K., Khan, R. U., Alkhalifah, A., & Ahmad, N. (2015). Urdu text classification using decision trees. *In 2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET)* (pp. 1-4). IEEE.

Kholladi, M. M. K. (2013). *Combinaison de classifieurs pour la reconnaissance de mots arabes manuscrits* (Doctoral dissertation, Université 20 aout 1955 de Skikda).

Khorsheed, M. S. (2002). Off-line Arabic character recognition–a review. *Pattern analysis & applications*, 5(1), 31-45.

Khoshnevisan, B., Bolandnazar, E., Barak, S., Shamshirband, S., Maghsoudlou, H., Altameem, T. A., & Gani, A. (2015). A clustering model based on an evolutionary algorithm for better energy use in crop production. *Stochastic environmental research and risk assessment*, 29(8), 1921-1935.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.

Kvarnhammar, A. M., & Cardell, L. O. (2012). Pattern-recognition receptors in human eosinophils. *Immunology*, 136(1), 11-20.

Lan, R. S., Yang, J. W., & Tang, Y. Y. (2009, July). A composite of central and ring projection. *In 2009 International Conference on Wavelet Analysis and Pattern Recognition* (pp. 200-204). IEEE.

Lawgali, A. (2015). A survey on arabic character recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(2), 401-426.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *In 2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8595-8598). IEEE.

Lee, J., & Kang, H. (2010). Flood fill mean shift: A robust segmentation algorithm. *International Journal of Control, Automation and Systems*, 8(6), 1313-1319.

Li, N., Kong, H., Ma, Y., Gong, G., & Huai, W. (2016). Human performance modeling for manufacturing based on an improved KNN algorithm. *The International Journal of Advanced Manufacturing Technology*, 84(1-4), 473-483.

Lorigo, L. M., & Govindaraju, V. (2006). Offline Arabic handwriting recognition: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 28(5), 712-724.

Lu, J., Liong, V. E., & Zhou, J. (2017). Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(8), 1979-1993.

Luqman, H., Mahmoud, S. A., & Awaida, S. (2015). Arabic and Farsi Font Recognition: Survey. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(01), 1553002.

Mahmoud, S. (2008). Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models. *Signal processing*, 88(4), 844-857.

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13), 2208-2217.

Mehta, S., Shen, X., Gou, J., & Niu, D. (2018). A New Nearest Centroid Neighbor Classifier Based on K Local Means Using Harmonic Mean Distance. *Information*, 9(9), 234.

Mesleh, A., Sharadqh, A., Al-Azzeh, J., Abu-Zaher, M., Al-Zabin, N., Jaber, T & Hasn, M. A. (2012). An optical character recognition. *Contemporary Engineering Sciences*, 5(11), 521-529.

Motawa, D., Amin, A., & Sabourin, R. (1997, August). Segmentation of Arabic cursive script. *In Proceedings of the Fourth International Conference on Document Analysis and Recognition* (Vol. 2, pp. 625-628). IEEE.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2016). Automatic text classification of ICD-10 related CoD from complex and free text forensic autopsy reports. *In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1055-1058). IEEE.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Al-Garadi, M. A. (2017). Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PloS one*, 12(2), e0170242.

Mustafa, W. A. (2017). A proposed optimum threshold level for document image binarization. *J. Adv. Res. Comput. Appl*, 7(1), 8-14.

Naz, S., Umar, A. I., Shirazi, S. H., Ahmed, S. B., Razzak, M. I., & Siddiqi, I. (2016). Segmentation techniques for recognition of Arabic-like scripts: A comprehensive survey. *Education and Information Technologies*, 21(5), 1225-1241.

Naz, S., Razzak, M. I., Hayat, K., Anwar, M. W., & Khan, S. Z. (2014). Challenges in baseline detection of Arabic script based languages. In *Intelligent Systems for Science and Information* (pp. 181-196). Springer, Cham.

Obaidullah, S. M., Goswami, C., Santosh, K. C., Das, N., Halder, C., & Roy, K. (2017). Separating Indic scripts with matra for effective handwritten script identification in multi-script documents. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(05), 1753003.

Ong, V., & Suhartono, D. (2016). Using K-Nearest Neighbor in Optical Character Recognition. ComTech: *Computer, Mathematics and Engineering Applications*, 7(1), 53-65.

Pal, U., Jayadevan, R., & Sharma, N. (2012). Handwriting recognition in indian regional scripts: a survey of offline techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1), 1.

Papa, J. P., Cappabianco, F. A., & Falcao, A. X. (2010, August). Optimizing optimum-path forest classification for huge datasets. *In 2010 20th International Conference on Pattern Recognition* (pp. 4162-4165). Ieee.

Parvez, M. T., & Mahmoud, S. A. (2013). Arabic handwriting recognition using structural and syntactic pattern attributes. *Pattern Recognition*, 46(1), 141-154.

Pattin, K. A., Greene, A. C., Altman, R. B., Cohen, K. B., Wethington, E., Görg, C & Moore, J. H. (2015). Training the next generation of quantitative biologists in the era of big data. *In Pacific Symposium on Biocomputing* (pp. 488-492).

Pechwitz, M., & Maergner, V. (2003). HMM based approach for handwritten Arabic word recognition using the IFN/ENIT-database. *In Seventh International Conference on Document Analysis and Recognition*. (pp. 890-894). IEEE.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4), 1432-1462.

Porro-Munoz, D., Duin, R. P., Orozco-Alzate, M., Talavera, I., & Londono-Bonilla, J. M. (2010). Classifying three-way seismic volcanic data by dissimilarity representation. *In 2010 20th International Conference on Pattern Recognition* (pp. 814-817). IEEE.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

Qin, H., Li, X., Yang, Z., & Shang, M. (2015, October). When underwater imagery analysis meets deep learning: a solution at the age of big visual data. In *OCEANS 2015-MTS/IEEE Washington* (pp. 1-5). IEEE.

Rahman, M. N., Esmailpour, A., & Zhao, J. (2016). Machine learning with big data an efficient electricity generation forecasting system. *Big Data Research*, 5, 9-15.

Razak, Z., Zulkiflee, K., Idris, M. Y. I., Tamil, E. M., Noor, M. N. M., Salleh, R., Yaacob, M. (2008). Off-line handwriting text line segmentation: A review. *International journal of computer science and network security*, 8(7), 12-20.

Ridzuan, F., Shirzad, Z., Azni, A. H., & Saudi, M. M. (2017). Classification for Quran Authentication Using Characters and Diacritics Hashed Values. *Advanced Science Letters*, 23(5), 4692-4695.

Rodriguez, J. A., & Perronnin, F. (2008). Local gradient histogram features for word spotting in unconstrained handwritten documents. *Proc. 1st ICFHR*, 7-12.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323-2326.

Roy, P. P., Bhunia, A. K., Das, A., Dey, P., & Pal, U. (2016). HMM-based Indic handwritten word recognition using zone segmentation. *Pattern Recognition*, 60, 1057-1075.

Roy, P. P., Bhunia, A. K., Das, A., Dey, P., & Pal, U. (2016). HMM-based Indic handwritten word recognition using zone segmentation. *Pattern Recognition*, 60, 1057-1075.

Rueda, L., & Herrera, M. (2008). Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space. *Pattern Recognition*, 41(10), 3138-3152.

Saabni, R., Asi, A., & El-Sana, J. (2014). Text line extraction for historical document images. *Pattern Recognition Letters*, 35, 23-33.

Sabbah, T., & Selamat, A. (2013). A framework for Quranic verses authenticity detection in online forum. In *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences* (pp. 6-11).

Sabbah, T., & Selamat, A. (2014, September). Support vector machine based approach for quranic words detection in online textual content. In *2014 8th. Malaysian Software Engineering Conference (MySEC)* (pp. 325-330).

Saber, Z., Sabri, A. Q. M., Kamsin, A., & Hakak, S. (2017). Efficient Approach to Segment Ligatures and Open Characters in Offline Arabic text. *Int. J. Comput. Commun. Instrum. Eng*, 4(1), 40-44.

Saeed, K., & Albakoor, M. (2009). Region growing based segmentation algorithm for typewritten and handwritten text recognition. *Applied Soft Computing*, 9(2), 608-617.

Sari, C., Akgül, C. B., & Sankur, B. (2013). Combination of gross shape features, fourier descriptors and multiscale distance matrix for leaf recognition. In *Proceedings ELMAR-2013* (pp. 23-26). IEEE.

Sari, T., Souici, L., & Sellami, M. (2002). Off-line handwritten Arabic character segmentation algorithm: ACSA. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition* (pp. 452-457).

Schiuma, G., Vuori, V., & Okkonen, J. (2012). Knowledge sharing motivational factors of using an intra-organizational social media platform. *Journal of knowledge management*, Vol. 16 No. 4, pp. 592-603.

Schuelke-Leech, B.-A., Barry, B., Muratori, M., & Yurkovich, B. (2015). Big Data issues and opportunities for electric utilities. *Renewable and Sustainable Energy Reviews*, 52, 937-947.

Shaikh, N. A., Mallah, G. A., & Shaikh, Z. A. (2009). Character segmentation of Sindhi, an Arabic style scripting language, using height profile vector. *Australian Journal of Basic and Applied Sciences*, 3(4), 4160-4169.

Shanthi, C., & Pappa, N. (2017). An artificial intelligence based improved classification of two-phase flow patterns with feature extracted from acquired images. *ISA transactions*, 68, 425-432.

Slimane, F., Kanoun, S., Alimi, A. M., Ingold, R., & Hennebert, J. (2010). Gaussian mixture models for arabic font recognition. *In 2010 20th International Conference on Pattern Recognition* (pp. 2174-2177). IEEE.

Slimane, F., Kanoun, S., Hennebert, J., Alimi, A. M., & Ingold, R. (2013). A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution. *Pattern Recognition Letters*, 34(2), 209-218.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

Spera, E., Tegolo, D., & Valenti, C. (2015). Segmentation and feature extraction in capillaroscopic videos. In *Proceedings of the 16th International Conference on Computer Systems and Technologies* (pp. 244-251). ACM.

Tagougui, N., Kherallah, M., & Alimi, A. M. (2013). Online Arabic handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(3), 209-226.

Tang, Y. Y., Cheng, H. D., & Suen, C. Y. (1991). Transformation-ring-projection (TRP) algorithm and its VLSI implementation. *International Journal of Pattern Recognition and Artificial Intelligence*, 5(01n02), 25-56.

Taylor, I., & Taylor, M. M. (2014). *Writing and Literacy in Chinese, Korean and Japanese: Revised edition* (Vol. 14): John Benjamins Publishing Company.

Terasawa, K., & Tanaka, Y. (2009, July). Slit style HOG feature for document image word spotting. In *2009 10th International Conference on Document Analysis and Recognition* (pp. 116-120). IEEE.

Vakil, M. I., Megherbi, D. B., & Malas, J. A. (2016, May). Optimized NCC-information theoretic metric for noisy wavelength band specific similarity measures. In *2016 IEEE Symposium on Technologies for Homeland Security (HST)* (pp. 1-6). IEEE.

Versteegh, K. (2014). *Arabic Language*: Edinburgh University Press.

Wakahara, T., Kimura, Y., & Tomono, A. (2001). Affine-invariant recognition of gray-scale characters using global affine transformation correlation. *IEEE transactions on pattern analysis and machine intelligence*, 23(4), 384-395.

Wan, M., Li, M., Yang, G., Gai, S., & Jin, Z. (2014). Feature extraction using two-dimensional maximum embedding difference. *Information Sciences*, 274, 55-69.

Xie, J., Zhang, L., You, J., & Shiu, S. (2015). Effective texture classification by texton encoding induced statistical features. *Pattern Recognition*, 48(2), 447-457.

Younis, K. S., & Alkhateeb, A. A. (2017). A new implementation of deep neural networks for optical character recognition and face recognition. *Proceedings of the new trends in information technology, Jordan*, 157-162.

Zanchettin, C., Bezerra, B. L. D., & Azevedo, W. W. (2012). A KNN-SVM hybrid model for cursive handwriting recognition. *In The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Zhao, W., Tang, S., & Dai, W. (2012). An improved kNN algorithm based on essential vector. *Elektronika ir Elektrotechnika*, 123(7), 119-122.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237, 350-361.