Contemporary Engineering Sciences, Vol. 5, 2012, no. 11, 521 - 529

An Optical Character Recognition

Abdelwadood Mesleh[1], Ahmed Sharadqh, Jamil Al-Azzeh,
MazenAbu-Zaher, Nawal Al-Zabin, Tasneem Jaber, Aroob Odeh
and Myssa'a Hasn

Computer Engineering Department, Faculty of Engineering Technology, Al-Balqa’
Applied University, Amman, Jordan
[1] e-mail: wadood@bau.edu.jo

Abstract

Arabic optical character recognition (OCR) is the process of converting images that contain Arabic text to a format that can be edited. In this work, a simple approach for Arabic OCR is presented, the proposed method deployed correlation and dynamic-size windowing to segment and to recognize Arabic characters. The proposed coherent template recognition process is characterized by the ability of recognizing Arabic characters with different sizes. Recognition results reveal the robustness of the proposed method.

Keywords: Arabic OCR, Arabic text recognition, Arabic recognition

1 Introduction

Humans recognize characters easily and they repeat the character recognition process thousands of times every day as they read papers or books. However, after many years of intensive investigation and research, the ultimate goal of developing an optical character recognition (OCR) system with the same reading capabilities as humans still remains unachieved. One of the main objectives of an OCR is to reach a 5 characters/second speed with a 99.9% recognition rate, with no errors. OCR is the process of converting an image representation of a document into an editable format. In the middle of the 1940s, the first character recognizers appeared and mainly focused on machine-printed text, and some of them dealt with handwritten text or symbols. In 1950s, commercial character recognizers were available for Latin languages. In 1980s, many structural and statistical methods were used in character recognition; some of those recognizers broke the character image into a set of lines and curves and basically focused on the shape recognition techniques without using any semantic information. After

1990, complex character recognition algorithms were developed; many recognizers used sophisticated methodologies such as neural networks, hidden Markov models and natural language processing techniques. Many applications such as reading postal address off envelopes, reading customer filled forms, archiving and retrieving text and digitizing libraries benefit from OCR systems. OCRs are divided into two major categories: typewritten and handwritten. Typewritten OCR systems recognize a document that has been previously typed and scanned prior to recognition progress. On the other hand, handwritten OCR systems recognize a text that has been written by a human. Comparing to handwritten OCR systems, Typewritten OCR systems are usually easier to design and the recognition rate achieved for typewritten recognition systems is more than the handwritten. OCRs are further categorized to offline and online recognition systems. In offline OCR systems, the image of the typewritten or the handwritten text is acquired through scanning. The image then is read by the OCR system and is analyzed for recognition. In online OCR systems, input of the OCR system is an image of a handwritten text which is usually acquired using cell phone or a portable personal computer. A large number of OCR research papers have been published on Latin, Chinese and Japanese characters. In fact, most Latin OCR algorithms assume that individual characters can be isolated, however, this is not true for languages with cursive scripts such as Arabic, as a result, little work has been published on the Arabic OCR and the progress of Arabic character recognition is still far behind the progress achieved in Latin and other languages.

Arabic OCR, the associated text recognition technologies, the characteristics of the Arabic language with respect to OCR and discusses related research on the different phases of text recognition are surveys in (Al-Muhtaseb & Qahwaji, 2011), moreover, the available databases for Arabic OCR research, the available commercial softwares, the challenges related to Arabic OCR and possible future trends are all discussed. In this paper, a new simple typewritten, offline character recognition for Arabic language is presented. This rest of this paper is organized as follows. Section 2 describes some of the related work of the Arabic OCR systems, sections 3 describes the proposed Arabic OCR algorithm, section 4 shows experimental results, and finally, the conclusion and the future work are presented in sections 5 and 6 respectively.

2 Related work of the Arabic OCR systems

In 2009, a distributed Arabic OCR based on the dynamic time warping algorithm is proposed, results shown that grid computing framework speeds up the Arabic OCR. AbdelRaouf et al. (2010) presented a comprehensive study and analysis of a multi-modal Arabic corpus that is suitable for use in OCR development. A distance classifier and artificial neural network classifier were used to discriminate between the different characters in an Arabic automatic license plate recognition system (Alginahi, 2011; Maglad, 2012). Dreuw et al.

(2012) presented a hidden Markov model based RWTH[1] OCR system that represents a unique framework for large vocabulary OCR, their proposed OCR system works for both handwriting and machine-printed Arabic texts. Oujaoura et al.(2012) developed an OCR system of isolated Arabic printed characters and used Zernike moments, invariant moments and Walsh Transformation in feature extraction phase and used artificial neural networks in as a classifier. In (Abulnaja & Batawi, 2012), an N-version programming technique, a fault-tolerant technique, is used to improve the accuracy of Arabic OCR. Some other recent work paid attention to the feature selection process that may enhance the performance of Arabic OCR systems: Moussa et al. (2010) used texture analysis to extract global features to reduce the processing difficulties in a recognition system and to make the Arabic printed multi-font recognition successful. Zahedi & Eslami (2011) deployed a scale invariant feature transform method to extract a set of features in Farsi and Arabic language OCR systems. Other feature selection approaches that enhance Arabic OCR are presented in (Bahgat et al., 2012). Comprehensive reviews of other work in Arabic OCR are discussed in (Khorsheed, 2002; AL-Shatnawi et al., 2011).

3 The proposed Arabic OCR algorithm

Arabic is the official language of over twenty Arab countries which stretch from Morocco to Iraq, it is the religious language of all Muslims of more than one billion Muslims spread all over the world and it is the language of the Quran (the sacred book of Islam). Arabic language is a Semitic language and most of its words are built up from roots by following certain fixed patterns and adding infixes, prefixes and suffixes. Arabic is an old language, and what is now known as Classical Arabic was standardized around fourteen centuries ago. The modern form of Arabic is called Modern Standard Arabic (MSA) and it is the form used in all Arabic-speaking countries in publications and media. The alphabet set used to write this language is the Arabic alphabet, (see Table 1). There is a number of languages that use the Arabic alphabet, such as Persian, Kurdi and Jawi. The characteristics of Arabic script make Arabic OCR more challenging, Arabic characters characterized by the following challenges (AL-Shatnawi et al., 2011): (i) The Arabic script is cursive. (ii) Arabic characters may have different shapes in different positions of a word; these shapes are grouped in 100 character shapes that present a lot of similarities (see Table 1). (iii) Most Arabic letters have one, two, or three dots. (iv) An Arabic word is composed of sub-word(s). (v) The Arabic font is written-read from right to left. (vi) Arabic characters are connected.

[1] The RWTH OCR system is based on a speech recognition framework RWTH-ASR-The RWTH Aachen University Speech Recognition System, and can be obtained from: http://www-i6.informatik.rwth-aachen.de/rwth-ocr/

(vii) There are two kinds of spaces between words and within a word introduced by characters that have no middle shape, moreover, Arabic characters are written following a writing line called herein "baseline" which is about 2/3 down the main body of the characters.

Table 1. Arabic characters.

Character Name		Isolated	Connected		
			Start	Mid	End
Alif	ألف	ا	ا	ـا	ـا
Baa	باء	ب	بـ	ـبـ	ـب
Taa	تاء	ت	تـ	ـتـ	ـت
Thaa	ثاء	ث	ثـ	ـثـ	ـث
Jeem	جيم	ج	جـ	ـجـ	ـج
Haa	حاء	ح	حـ	ـحـ	ـح
Khaa	خاء	خ	خـ	ـخـ	ـخ
Daal	دال	د	د	ـد	ـد
Thaal	ذال	ذ	ذ	ـذ	ـذ
Raa	راي	ر	ر	ـر	ـر
Zaay	زاي	ز	ز	ـز	ـز
Seen	سين	س	سـ	ـسـ	ـس
Sheen	شين	ش	شـ	ـشـ	ـش
Saad	صاد	ص	صـ	ـصـ	ـص
Saad	صاد	ص	صـ	ـصـ	ـص
Dhaad	ضاد	ض	ضـ	ـضـ	ـض
Ttaa	طاء	ط	طـ	ـطـ	ـط
Dthaa	ظاء	ظ	ظـ	ـظـ	ـظ
Ain	عين	ع	عـ	ـعـ	ـع
Ghen	غين	غ	غـ	ـغـ	ـغ
Faa	فاء	ف	فـ	ـفـ	ـف
Qaf	قاف	ق	قـ	ـقـ	ـق
Kaf	كاف	ك	كـ	ـكـ	ـك
Lam	لام	ل	لـ	ـلـ	ـل
Mem	ميم	م	مـ	ـمـ	ـم
Noon	نون	ن	نـ	ـنـ	ـن
Haa	هاء	ه	هـ	ـهـ	ـه
Wow	واو	و	و	ـو	ـو
Yaa	ياء	ي	يـ	ـيـ	ـي

(viii) Arabic words may contain ligatures and overlapping. (ix) Some characters are overlapped in some Arabic words, overlapping occurs whenever two or more characters overlap each other. (x) Some Arabic words contain diacritics. An Arabic OCR system may contain all or part of the following steps (Khorsheed, 2002; AL-Shatnawi et al., 2011): Image Acquisition, Image preprocessing, Line segmentation, Word segmentation, Character segmentation and Recognition. The recognition process can be divided into two main approaches: a "Word-by-word" approach that works like a dictionary and needs a large database of words and a "Character-by-character" approach that identifies characters to form words. It is known that the main problem of the "word-by-word" approach does not work well with cursive script, such as Arabic, and it is difficult to isolate characters using "“Character-by-character" approach where the shape of the shape of Arabic characters is context sensitive. Thus, most of the currently proposed OCR algorithms fail to exhibit an appropriate recognition rate when facing cursive documents. In this paper, the proposed OCR algorithm combines the word segmentation, character segmentation and recognition steps in a coherent template process step. The proposed AOCR algorithm follows the following three main steps (See Figure 1):

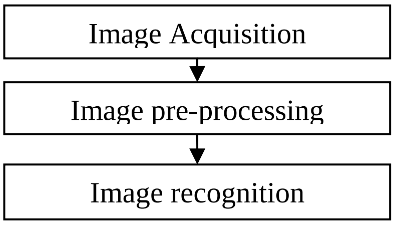


Figure 1. Proposed Arabic OCR algorithm.

Step1: *Image acquisition*: the proposed AOCR starts with image acquisition process (see Figure 2) that scans the Arabic text using a 300 dpi scanner, the scanned image is saved in a .bmp image file.
Step 2: *Image pre-processing*: the image is filtered using a median filter that removes noise, the image is converted from RGB to gray scale image, and then it is converted to binary image. It should be noted that not removing the noise may lead to incorrect results in the recognition process. The area outside text boundaries is removed using a clipping process and finally, the binary image is resized to a common defined size as shown in Figure 3.
Step 3: *Image recognition*: this step equivalent to the mentioned line segmentation, word segmentation, character segmentation and recognition steps.
(i) The binary image is segmented (Zeki et al., 2011) into lines of text; Figure 4 shows the two lines of text that are segmented from the binary image in Figure 3.

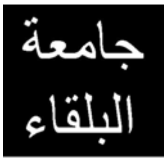


Figure 2. Original image.

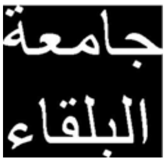
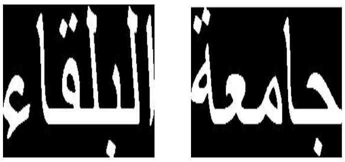


Figure 3: Binary clipped image.



Second line First line

Figure 4: Line segmentation.

(ii) For each segmented and cropped line, the "maximum vertical coordinate" (see the upper red line in Figure 5) and "base line coordinate" (see the lower red line in Figure 5) are calculated, and then the font size is calculated by subtracting the "base line coordinate" from the "maximum vertical coordinate", finally, the size of the dynamic-size window is determined to fit the size of the templates of the previously collected Arabic characters (templates of all standard Arabic characters

with different shapes, sizes and types that are collected from Alrai news paper[2]).
(iii) The proposed Arabic OCR algorithm shall increase the width of the dynamic-sized window until some valid character is matched, on the other hand, the height is determined by the cropping function.
(iv) As Arabic characters are cursive, the dynamic-sized window may contain a full character of a portion of a character, to solve this problem, the proposed algorithm need to take a decision whether to accept or reject the character that is contained in the dynamic-sized window. If it is rejected, the width of the window is increased by some constant value (C) until a valid character is found (empirically, C is set to 5 pixels). It should be noted that the width of the window is initially set to a constant value=font size/10 and the search for candidate characters starts from right of each segmented and clipped line.
(v) In some other cases, the dynamic-sized window may contain a portion of the successor or previous character, the proposed Arabic OCR algorithm removes remnants of the other Arabic characters (see Figure 6). It should be noted that image comparison uses correlation to determine whether the candidate character is valid or not. If correlation is bigger than 0.8, the character is a valid Arabic character and it is written in note pad. However, if the correlation is less than 0.8, the proposed Arabic OCR algorithm increases the width of dynamic-sized window by C (see Figure 7).

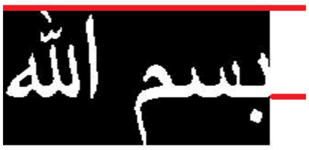


Figure 5: Font size calculation.

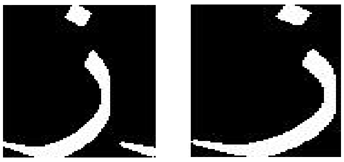


Image with remnants

Character image

Figure 6: Remnants filtering.

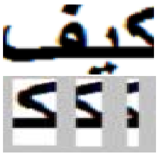


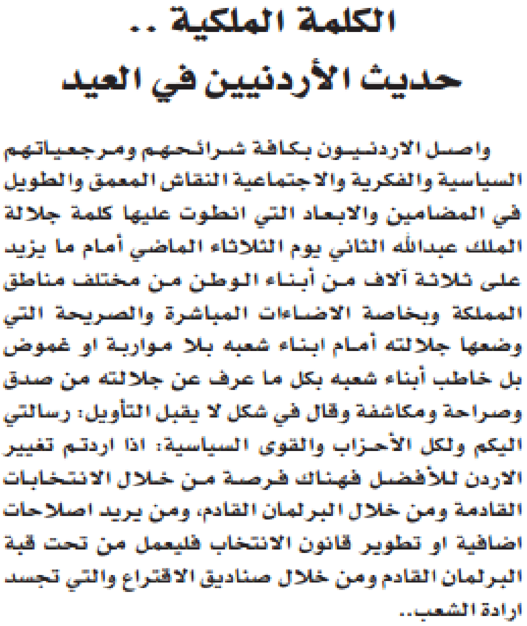
Figure 7: Dynamic-sized window width-sizing

4. Experimental results

To illustrate the accuracy of the proposed Arabic OCR algorithm, the performance was tested using articles from Alrai news paper on 20/11/2011 and 30/11/2012, Figure 8 shows a sample document retrieved from Alrai news paper on Oct, 20, 2012. They were scanned using HP scanjet 2400 scanner and the images were then filtered, binarized, clipped and resized. Lines of text were then extracted from the images. The font size was identified; segmentation was performed on each line to segment characters taking in consideration the characteristics of Arabic scripts such as overlapping and the recognition of Arabic text with different sizes. MATLAB (R2012.a/64-bit) is used to implement the

[2] A Jordanian news paper: http://www.alrai.com/

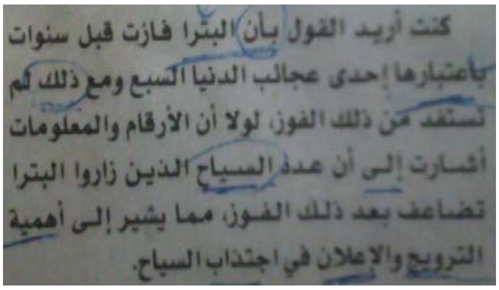
proposed Arabic OCR algorithm on an HP Pavilion g6 (Intel(R) Core(TM) i5 CPU M480 @ 2.67GHz with 3.00 GB RAM) machine running a 64-bit operating system -MS Windows 7. The recognition accuracy was 96.5% due to novel online character segmentation and recognition method that handled the uniqueness of Arabic script (i.e., its cursive and overlapping natures). The templates of all typewritten Arabic characters were manually collected and processed from many articles in Alrai news paper; these templates contain Arabic characters in all possible shapes. To test the robustness of the proposed Arabic OCR algorithm, a noisy paragraph is randomly selected from a published article in Alrai newspaper on 20/11/2011 (see Figure 9). The recognition result is shown in Figure 10. It is obvious that the proposed Arabic OCR algorithm recognized it correctly.



الكلمة الملكية ..
حديث الأردنيين في العيد

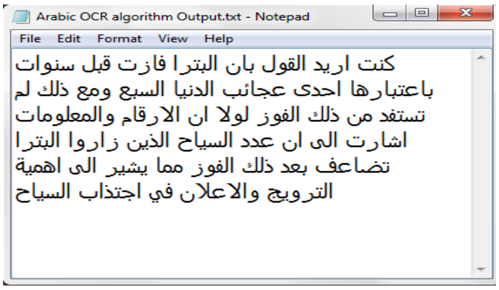
واصـل الاردنـيـون بـكـافـة شـرائـحـهـم ومـرجـعـيـاتـهـم
السياسية والفكرية والاجتماعية النقاش المعمق والطويل
في المضامين والابعاد التي انطوت عليها كلمة جلالة
الملك عبدالله الثاني يوم الثلاثاء الماضي أمام ما يزيد
على ثلاثة آلاف من أبناء الوطن من مختلف مناطق
المملكة وبخاصة الاضاءات المباشرة والصريحة التي
وضعها جلالته أمام ابناء شعبه بلا مواربة او غموض
بل خاطب أبناء شعبه بكل ما عرف عن جلالته من صدق
وصراحة ومكاشفة وقال في شكل لا يقبل التأويل: رسالتي
اليكم ولكل الأحزاب والقوى السياسية: اذا اردتم تغيير
الاردن للأفضل فهناك فرصة من خلال الانتخابات
القادمة ومن خلال البرلمان القادم، ومن يريد اصلاحات
اضافية او تطوير قانون الانتخاب فليعمل من تحت قبة
البرلمان القادم ومن خلال صناديق الاقتراع والتي تجسد
ارادة الشعب..

Figure 8: Test sample text image.



كنت أريد القول بأن البترا فازت قبل سنوات
باعتبارها إحدى عجائب الدنيا السبع ومع ذلك لم
تستفد من ذلك الفوز لولا أن الأرقام والمعلومات
أشارت إلى أن عدد السياح الذين زاروا البترا
تضاعف بعد ذلك الفوز مما يشير إلى أهمية
الترويج والإعلان في اجتذاب السياح.

Figure 9: A noisy scanned text image.



Arabic OCR algorithm Output.txt - Notepad
File Edit Format View Help

كنت اريد القول بان البترا فازت قبل سنوات
باعتبارها احدى عجائب الدنيا السبع ومع ذلك لم
تستفد من ذلك الفوز لولا ان الارقام والمعلومات
اشارت الى ان عدد السياح الذين زاروا البترا
تضاعف بعد ذلك الفوز مما يشير الى اهمية
الترويج والاعلان في اجتذاب السياح

Figure 10: Recognition results.

5. Conclusion and future work

In this paper, a template based Arabic OCR algorithm is proposed, the proposed algorithm deploys correlation and dynamic-size windowing to segment and to recognize Arabic characters and it recognizes Arabic characters with different sizes. It achieves 96% recognition accuracy.

Handling Arabic text with different orientation and using artificial intelligent methods instead of correlation are left for future work.

References

[1] A. AL-Shatnawi, S. AL-Salaimeh, F. AL-Zawaideh and O. Khairuddin, Offline Arabic Text Recognition-An Overview, World of Computer Science and Information Technology Journal (WCSIT), 1(2011), 184-192.

[2] Y. Alginahi, Automatic Arabic License Plate Recognition, International Journal of Computer and Electrical Engineering, 3(2011), 454-460.

[3] M. Khorsheed, Off-Line Arabic Character Recognition – A Review, Pattern Analysis & Applications, 5(2002), 31-45.

[4] AbdelRaouf, C. Higgins, T. Pridmore and M. Khalil, Building a multi-modal Arabic corpus (MMAC), International Journal on Document Analysis and Recognition, 13(2010), 285-302.

[5] M. Khemakhem and A. Belghith, Towards A Distributed Arabic OCR Based on the DTW Algorithm: Performance Analysis, The International Arab Journal of Information Technology, 6(2009), 153-161.

[6] P. Dreuw, D. Rybach, G. Heigold and H. Ney, RWTH OCR: A Large Vocabulary Optical Character Recognition System for Arabic Scripts, in V. Märgner and H. El Abed (ed.), Guide to OCR for Arabic Scripts Chapter, Part II: Recognition, 2012, 215-254, Springer, London, UK.

[7] S. Moussa, A. Zahour, A. Benabdelhafid and A. Alimi, New features using fractal multi-dimensions for generalized Arabic font recognition, Pattern Recognition Letters, 31(2010), 361-371.

[8] M. Zahedi and S. Eslami, Farsi/Arabic Optical Font Recognition Using SIFT Features, Procedia Computer Science, 3(2011), 1055-1059.

[9] M. Oujaoura, R. El Ayachi, M. Fakir, B. Bouikhalene and B. Minaoui, "Zernike moments and neural networks for recognition of isolated Arabic characters," International Journal of Computer Engineering Science, 2(2012), 17-25.

[10] O. Abulnaja and Y. Batawi, Improving Arabic Optical Character Recognition Accuracy Using N-Version Programming Technique, Canadian Journal on Image Processing and Computer Vision, 3(2012), 44-46.

[11] H. Al-Muhtaseb and R. Qahwaji, Arabic Optical Character Recognition: Recent Trends and Future Directions, Applied Signal and Image Processing: Multidisciplinary Advancements, ed. R. Qahwaji, R. Green and E. Hines, 2011, 324-346.

[12] Zeki, M. Zakaria and C. Liong, Segmentation of Arabic Characters: A Comprehensive Survey, International Journal of Technology Diffusion, 2(2011), 48-82.

[13] S. Bahgat, S. Ghomiemy, S. Aljahdali and M. Alotaibi, A Proposed Hybrid Technique for Recognizing Arabic Characters, International Journal of Advanced Research in Artificial Intelligence, 1(2012), 35-43.

[14] K. Maglad, A Vehicle License Plate Detection and Recognition System, Journal of Computer Science, 8(2012), 310-315.

Received: September, 2012