Pattern Recognition 47 (2014) 672–684

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Arabic font recognition based on diacritics features

Mohammed Lutf [a], Xinge You [a,*], Yiu-ming Cheung [b], C.L. Philip Chen [a,c]

[a] Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China
[b] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[c] Faculty of Science of Technology, University of Macau, China

ARTICLE INFO

Article history:
Received 7 November 2012
Received in revised form
17 July 2013
Accepted 21 July 2013
Available online 3 August 2013

Keywords:
Font recognition
Arabic diacritics
Composite of central and ring projection

ABSTRACT

Many methods have been proposed for Arabic font recognition, but none of them has considered the specialty of the Arabic writing system. Most of these methods are either general pattern recognition approaches or application of other methods which have been developed for languages other than Arabic. Therefore, this paper is the first attempt to present an alternative method for Arabic font recognition based on diacritics. It presents the diacritics as the thumb of Arabic fonts which can be used individually to identify and recognize the font type. Diacritics are the marks and strokes which have been added to the original Arabic alphabet. Though they are the smallest regions in the Arabic script, with today technology it is very easy to get a high resolution image with a very low cost. In this kind of images, the diacritics can reveal very useful information about the font type. In this study, two algorithms for diacritics segmentation have been developed, namely flood-fill based and clustering based algorithm. The experiments conducted proved that our approach can achieve an average recognition rate of 98.73% on a typical database that contains 10 of the most popular Arabic fonts. Compared with existing methods, our approach has the minimum computation cost and it can be integrated with OCR systems very easily. Moreover, it could recognize the font type regardless of the amount of the input data since five diacritics, which in most cases can be found in only one word, are enough for font recognition.

1. Introduction

Font recognition is a fundamental issue in the identification and analysis of documents. In the past this task was considered highly demanding on computer hardware. The OCR techniques are making great successes in commercial software, but the possibility of increasing the efficiency of the OCR system can be guaranteed by taking font type into account. Automatic document processing (ADP) techniques tackle font recognition on the basis of two main aspects. First, it generalizes the font type for all characters in the document. The use of this approach only enables us to reduce the number of alternative forms of each class of a font family. This clearly leads to the recognition of only one kind of font. The second aspect that should be considered in ADP techniques is the identification of the font types used within the document, which is usually neglected in spite of its importance [1].

Different Optical Font Recognition (OFR) methods have been successfully applied in many languages except Arabic, because these methods always fail to accord with the characteristics of the Arabic writing system. Arabic is an alphabetic language written in a cursive way and this is not only in handwriting but also in machine typing characters. It is something between Latin and Chinese. In Latin, each word consists of letters separated from each other and in Chinese each character consists of strokes connected with each other to build one block representing a word. In Arabic, however, the word consists of one or more subwords and each subword may have one or more letters connected with each other, which makes working with an Arabic text a very challenging task. This explains why these approaches are successful with most languages but not with Arabic.

The most common approaches toward font recognition are font recognition based on typographical features [2–4] and font recognition using textural features [5–8]. It has been reported that the second approach is more efficient than the former one [9]. However, solving the problem of font recognition is still just a small task in OCR or ADP system. Most of the existing OFR methods are full pattern recognition systems; they have their own processing blocks starting from reading the image to the final step which is the recognition of the font class. When one of these types of OFR systems is added to OCR or ADP system, it increases the computational cost and mainly reduces the performance of the system especially when it is used in real time applications, for example, using OCR system to read the road signs in an autonomous car. Therefore, the speed and the possibility of integrating

* Corresponding author. Tel.: +86 2787544817; fax: +86 2787544823.
E-mail addresses: Mohammed.lutf@gmail.com (M. Lutf), you1231cncn@gmail.com, youxg@mail.hust.edu.cn, XingeYou@hust.edu.cn (X. You), ymc@comp.hkbu.edu.hk (Y.-m. Cheung), philip.Chen@ieee.org (C.L. Philip Chen).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.015

the OFR with other systems through sharing most of the processing blocks is a critical need. Another drawback in the currently proposed methods is that when identifying the font type, the features depend on the whole given text image and it is always preceded by a preprocessing step. This fact makes the font recognition even more costly. The study reported in this paper addresses the problems discussed above and proposes instead an Arabic font recognition system which can easily be integrated with OCR system by sharing all its preprocessing blocks. At the same time, the features are extracted only from the diacritics which are very easy to be segmented.

Over the last two decades, a few approaches were proposed for Arabic font recognition. It can be classified into two main categorizes: segmentation-free and segmentation-based approaches. Segmentation-free approaches intend to recognize the font by extracting the texture features globally from a predefine text entity, it could be a whole image, a text block or a text line. Also, the texture features extraction algorithm is always a very well known algorithm. Systems based on this approach differ only on the type of the text entity and the algorithm used for extracting the texture features.

In [10,11], Silmane et al. used Gaussian mixture model to estimate the font category likelihoods in a word images. It was the first attempt to evaluate Arabic OFR on a publicly available database(APTI), but due to the database limitation, this approach can be used only with already segmented word images. Imani et al. [12] proposed the use of wavelet to extract features from a 128×128 text block. In this approach, most of the training dataset was labeled using a learning algorithm which may produce wrong labeled data. As a result, this will reduce the final recognition rate. In [13], Bataineh et al. used Gray-Level Co-occurrence Matrix (GLCM) features extracted from the edges of a 512×512 text block. The text block was created by removing the spaces between words and lines, but multi-size words are not normalized which will produce different edge features for the same word written by the same font. Pourasad et al. [14] proposed the use of holes of letters and text line horizontal projection profile for Farsi font recognition. This approach will fail with fonts contain no holes and also will fail with the skewness text. Khosravi and Kabir [15] proposed the use of Sobel and Roberts gradients in 16 directions to extract the texture features form a 128×128 text block for Farsi font. This approach is flexible regarding the size of the input text block, but this is also a problem because the text block is processed without normalization and the measurements depend on pixels which make this approach fail with high resolution or fewer word input text. In [16], Ben Moussa et al. demonstrated how we can use a combination of BCD (box counting dimension) and DCD (dilation counting dimension) techniques to obtain features from text paragraphs, accuracy rate of about 98% has obtained. Abuhaiba [17] proposed the use of horizontal and vertical projection profiles, Walsh coefficients, invariant moments, geometrical features for Arabic OFR using word level images. The features in this approach are extracted from common words, and the author assumes that at least the whole paragraph is written using the same font. The problem with this approach is that the segmentation algorithm of the common words during the test process may filter out all the input text. Zramdini and Ingold [3] proposed the use of scale invariant feature transform (SIFT) with 128×128 text block. The authors claim to reach a recognition rate of 100% , but the computational cost is very high especially when using big database.

For segmentation-based approach, only one method was proposed by Abuhaiba [18]. This approach segments each word into symbols then creates a template for each symbol. The problem with this approach is that it is based on segmenting the individual characters in each word which is the most complicated problem in Arabic text. The method proposed in this work which uses the word height for symbol segmentation is not valid with the majority of Arabic fonts and it works only under ideal conditions.

In addition, all techniques used in these approaches are general techniques which can be applied to any texture analysis or recognition problem. It does not have any specific treatment for Arabic text where a simple modification or direct application of these techniques may not solve the problem of Arabic font recognition. And the absence of commercial products for Arabic OFR is an evidence.

Although the texture features approaches are robust to noisy and low resolution text images, it is reliable only with uniform and homogeneous text block where all words have the same font which is not always the case. So, to take the advantages of texture features robustness and to overcome the complexity of Arabic character segmentation, we propose a novel method[1] for Arabic font recognition using diacritics-based rotation invariant features with a low computational cost. Diacritics are not connected to each other nor to the text body which makes the diacritics very easy to be segmented. Two efficient algorithms have been developed for diacritics segmentation which are the main contribution in our method. Besides font recognition, we also address the computation simplification and font recognition of irregular text like skewness text lines, multi-font formats, and multi-language text image.

Diacritics are the most common shapes that appear in any Arabic text. Unlike the normal characters' shapes, where some of them may not be found at all; it is very easy to find hundreds of diacritics in only one page of a text. The dots diacritics, for instant, are shared by many characters and the same vowel diacritic can be attached to almost all characters. Therefore, the fact that our recognition system is based mainly on diacritics allows us to ensure getting sufficient information from any input image even if it contains only few words. Thus, our focus on diacritics does not mean that normal characters are useless; it is similar to identifying a person using only his fingerprint. If we consider the retina, face, voice, DNA and many other biometrics, we will have more discriminatory information; but as long as the fingerprint gets the task done, there is no need to include other factors. The same thing is applied to diacritics.

The rest of this paper is organized as follows. In Section 2, we introduce Arabic diacritics in detail. In Section 3, we introduce our proposed method. In Section 4, the experimental results are given. Finally, Section 5 concludes the paper and gives some perspectives of future work.

2. Arabic diacritics

Arabic is a widely used alphabetic writing system in the world [20], and it has 28 basic letters. The alphabet was first used to write texts in Arabic, most notably the Qur'an, the holy book of Islam as shown in Fig. 1. With the spread of Islam, it came to be used to write many languages like, at various times, Urdu, Pashto, Uyghur (in China), Ottoman Turkish and Spanish (in Western Europe) [21]. To accommodate the needs of these languages, new letters and symbols were added to the original alphabet. 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 to

1. Distinguish between letters of the same or similar shape.
2. Indicate sounds (vowels and tones) that are not conveyed by the basic alphabet.

[1] This work is an extension to our conference paper [19].