Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 65 (2015) 556 – 561

Procedia Computer Science

International Conference on Communication, Management and Information Technology (ICCMIT 2015)

# A Database for Arabic Handwritten Character Recognition

Jawad H AlKhateeb*

*Department Of Computer Science, College of Computer Science and Engnieering,Taibah University, KSA*

**Abstract**

This paper proposes an image database for Arabic handwritten character recognition (AHCR). In this paper, the Arabic handwritten images character database written by multi writers is proposed. This database is eligible for Arabic handwritten recognition research. The database contains the digital images of the Arabic alphabets written by 100 native Arabic writers. Each writer writes the Arabic letter 10 times on a form. All the forms were scanned using a high quality scanner. Earlier, all the Arabic characters were cropped form the forms. Therefore, the database contains 28000 images. These images were divided into two sets; 80% for training and 20% or testing. This database base will be freely available



Peer-review under responsibility of Universal Society for Applied Research

*Keywords:* Arabic Offline Handwritten Recognition, Pre-Processing, AHCR Database.

## 1. Introduction

A large number of research has been conducted for the recognition of Latin, Chinese, and Japanese text. On the other hand, relatively little research has been done on Arabic text. This is due to the complexity of Arabic text and to limited Arabic databases. Recognition of Arabic text is at the early stage compared to the methods for recognition of Latin, Chinese, and Japanese text. In addition, there is a major challenge in the Arabic writing recognition systems, which comes from the cursive nature of the data. Recognition of Arabic handwritten text is a difficult task. This

* Corresponding author. Tel.: +966569749207
*E-mail address:* jkhateeb@taibahu.edu.sa


Peer-review under responsibility of Universal Society for Applied Research
doi:10.1016/j.procs.2015.09.130

difficulty comes from many factors such as the Arabic writing mechanism which is cursive, the writer style, the pen, and other factors [1][2].

Many domains for Arabic handwritten recognition can be classified as, character recognition, office automation, cheques verification, mail sorting, and a large variety of banking, business as well as natural human-computer interaction[3]. In general, the Arabic handwritten task is divided into two main systems. First, the on-line based system where the process of writing is being traced by the computer. Hence the strength and sequential order of each segment when it is written can be recorded for recognition. Second, the off-line based systems in which, the digital image is available only. The off-line based system is more difficult [3][4][5].

The work for Arabic script recognition has started more than three decades ago. Al-Muallim and Yamaguchi [4] proposed a structural recognition technique for Arabic handwritten words which were segmented into strokes. The strokes were classified and combined into characters according to their features. However, their system showed a failure in most cases due to incorrect segmentation of words. Amin and Alsadoun [5] proposed techniques using binary tree to segment printed Arabic text into characters. Amin and Alsadoun [6] proposed recognition of hand printed Arabic characters using neural network. Abuhaiba [7] dealt with some problems in the processing of binary images of handwritten text documents, such as extracting lines from pages, which is found to be powerful and suitable for variable handwriting. Abuhaiba et al. [8] introduced a novel offline cursive Arabic script recognition system to recognize offline handwritten cursive script having high variability based on segmentation based system. In their system, a single component strokes were extracted. Khorsheed [9] presented a new method on off-line recognition of handwritten Arabic script, in which segmentation into characters is not required. The method decomposed the skeleton of the word into an observation sequence, and then a single hidden Markov model (HMM) with structural features is employed for classification. HMM is also used in Alma'adeed et al [10] for unconstrained Arabic handwritten word recognition. In Alma'adeed [11], a complete scheme for unconstrained Arabic handwritten word recognition based on a neural network is proposed.

Any recognition system ideally needs a large database to train and test the system. Real data from banks or the post code are confidential and inaccessible for non-commercial research. Although some work was conducted in Arabic handwritten digits, but generally they had small databases of their own or the presented results on databases which were unavailable to the public. Consequently, there was no benchmark to compare the results obtained by researches. The ADBase database is available for free, is very important in this context as it has been used as a standard test set in such a context [12].

El-sherif and Abdleazeem [12] released an Arabic handwritten digit database (ADBase) which is composed of 70,000 digits written by 700 writers. Each writer wrote each digit (from 0-9) ten times. To ensure including different writing styles, the database was gathered from different institutions: Colleges of Engineering and Law, School of Medicine, the Open University (whose students span a wide range of ages), a high school, and a governmental institution. Forms were scanned with 300 dpi resolution then digits are automatically extracted, categorized, and bounded by bounding boxes. The scanner was adjusted to produce binary images directly. Some noisy and corrupted digit images were edited manually. The database is divided into two sets: training and testing set. The training set includes 60,000 digits to 6,000 images per class, and the testing set includes 10,000 digits to 1000 images per class. The ADBase is available for free (http://datacenter.aucegypt.edu/shazeem/ ) for researchers.

A standard database of Arabic handwritten images is required to design any recognition system. Arabic handwritten recognition systems lack the standard databases since most of the Arabic handwritten research is conducted on private database. In this paper, we propose an image database for Arabic handwritten character recognition (AHCR).

## 2. Arabic Writing

The Arabic alphabet consists of 28 letters, and writing is written from right to left in a cursive manner.The Arabic alphabet is used for writing different languages such as Persian, Urdu, and Jawi. Each Arabic letter has either two or four shapes depending on its position in the text. The shape of a letter changes with its position, which may be at start, middle or end of a word, or alone [3][6]. Although generally Arabic is cursive, there are some non-cursive letters. There are 22 cursive letters with four different shapes and 6 non-cursive letters with only two shapes corresponding to the alone and end positions. Table 1 shows each shape for each letter. For example letter Ayn (ع) has the following shapes: عـ at start, ـعـ at middle, ـع at end, and ع when alone. Moreover, in order to distinguish some characters from