Chapter 16

Segmentation of Arabic Characters: A Comprehensive Survey

Ahmed M. Zeki
University of Bahrain, Bahrain

Mohamad S. Zakaria
Universiti Kebangsaan Malaysia, Malaysia

Choong-Yeun Liong
Universiti Kebangsaan Malaysia, Malaysia

ABSTRACT

The cursive nature of Arabic writing is the main challenge to Arabic Optical Character Recognition developer. Methods to segment Arabic words into characters have been proposed. This paper provides a comprehensive review of the methods proposed by researchers to segment Arabic characters. The segmentation methods are categorized into nine different methods based on techniques used. The advantages and drawbacks of each are presented and discussed. Most researchers did not report the segmentation accuracy in their research; instead, they reported the overall recognition rate which did not reflect the influence of each sub-stage on the final recognition rate. The size of the training/testing data was not large enough to be generalized. The field of Arabic Character Recognition needs a standard set of test documents in both image and character formats, together with the ground truth and a set of performance evaluation tools, which would enable comparing the performance of different algorithms. As each method has its strengths, a hybrid segmentation approach is a promising method. The paper concludes that there is still no perfect segmentation method for ACR and much opportunity for research in this area.

1. INTRODUCTION

The Arabic language is one of the most structured and served languages. It comes as the fifth of the most used languages (as a first language) after Chinese, Hindi, Spanish and English. It is spoken as a first language by nearly 350 million people around the globe, mainly in the Arab countries, which is about 5.5% of the world population (the world population is estimated at 6.44 billion in July 2005) (CIA, 2005). However, almost all Muslims (close to ¼ of the world population) can read Arabic script as it is the language of the Holy Qur'an.
The Arabic script evolved from a type of Aramaic, with the earliest known document dating from 512 AD. The Aramaic language has fewer consonants than Arabic (Burrow, 2004). The old Arabic was written without dots or diacritics. The dots were first introduced by Yahya bin Ya’mur (died around 746 AD) and Nasr bin Asim (died around 707 AD), students of Abu Al-Aswad Al-Du’ali (died around 688 AD) who introduced the diacritics to prevent the Qur’an from being misread by Muslims (Al-Fakhri, 1997). Figure 1 shows a sample of an old manuscript of a sentence written without dots or diacritics.

Due to the Islamic conquests, the use of Arabic language extended in the 7th and 8th centuries from India to the Atlantic Ocean (Al-Fakhri, 1997). Consequently, many other languages adopted the Arabic alphabet with some changes. Among those languages are Jawi, Urdu, Persian, Ottoman, Kashmiri, Punjabi, Dari, Pashto, Adighe, Baluchi, Ingush, Kazakh, Uzbek, Kyrgyz, Uygur, Sindhi, Lahnda, Hausa, Berber, Comorian, Mandinka, Wolof, Dargwa, and few others. Figure 2 shows samples of some of the above mentioned languages. However, it must be mentioned that some of those languages are currently using Latin characters, but in general, people can still read the Arabic script. It is also worth mentioning that the United Nation adopted Arabic in 1974 as its sixth official language (Strange, 1993).

Despite the fact that Arabic alphabets are used in many languages, Arabic Character Recognition (ACR) has not received enough interests from researchers. Little research progress has been achieved as compared to the one done on Latin or Chinese. It has almost only started in 1975 by Nazif (1975), while the earlier research efforts in Latin may be traced back to the middle of the 1940s. However, due to a lack of computing power, no significant work was performed until the 1980s. Recent years have shown a considerable increase in the number of research papers related to ACR.

The rest of this paper is organized as follows: the next section will introduce the Arabic Character Recognition in general. Section 3 will discuss the challenges faced by researchers attempting to segment Arabic characters. Section 4 reviews the methods used in segmenting the Arabic characters. Those methods are categorized under nine different categories based on the techniques used. The paper then ends with a discussion and conclusion.

2. ARABIC CHARACTER RECOGNITION

Character recognition is a major field in the area of pattern recognition which has been the subject of much research in the past four decades. The ultimate goal of any character recognition system is to simulate the human reading capabilities. A character recognition system is a program designed to convert a scanned document, which is seen by the computer as an image, into a text document that can be edited (Zeki & Ismail, 2002).

Character recognition systems can contribute tremendously to the advancement of the automation process and can improve the human computer interaction in many applications, including office automation, automatic mail routing, check verification, library archives, documents identifications,