Chapter 19
Arabic Optical Character Recognition:
Recent Trends and Future Directions

Husni Al-Muhtaseb
King Fahd University of Petroleum and Minerals, Saudi Arabia

Rami Qahwaji
University of Bradford, UK

ABSTRACT

Arabic text recognition is receiving more attentions from both Arabic and non-Arabic-speaking researchers. This chapter provides a general overview of the state-of-the-art in Arabic Optical Character Recognition (OCR) and the associated text recognition technology. It also investigates the characteristics of the Arabic language with respect to OCR and discusses related research on the different phases of text recognition including: pre-processing and text segmentation, common feature extraction techniques, classification methods and post-processing techniques. Moreover, the chapter discusses the available databases for Arabic OCR research and lists the available commercial Software. Finally, it explores the challenges related to Arabic OCR and discusses possible future trends.

INTRODUCTION

Arabic is the first language for more than 400 million people in the world. It is also used by more than 1 billion Muslims all over the world as a second language, for it is the language in which the Holy Qur'an was revealed. Arabic was added to the official languages of the United Nations in 1973 as the sixth language. The other five official languages (Chinese, English, French, Russian and Spanish) were chosen when the United Nations was founded. Also as has been reported by National Geographic (National Geographic, 2004), Arabic is expected to be one of the 5 major languages by 2050. Its importance is expected to rise, as English declines.

DOI: 10.4018/978-1-60960-477-6.ch019

Arabic is one of the Semitic languages. The Arabic script is being used/ had been used in other languages. Some of which are Hausa, Kashmiri, Kazak, Kurdish, Kyrghyz, Malay, Morisco, Pashto, Persian/Farsi, Punjabi, Sindhi, Tatar, Turkish, Uyghur, and Urdu (United Nations, 2006).

Arabic Optical Character Recognition (OCR) is an important and emerging application and research area. An OCR tool could be used to avoid retyping a scanned document or to convert the text images in the scanned document to an editable text. Such tool takes the scanned document as a picture and recognizes the text in the picture to make it available in text format.

Optical Arabic text recognition has received renewed extensive research after the recent successes in optical character recognition. Arabic text recognition, which was not researched as thoroughly as Latin, Chinese, or Japanese, is receiving more attentions from both Arabic and non-Arabic-speaking researchers.

Irrespective of the language under consideration, some typical applications of text recognition include: cheque verification, office automation, reading postal address, writer identification, and signature verification. Searching scanned documents available on the internet and searching Arabic historical manuscripts are also emerging applications. When Arabic is considered, there is real need to advance these applications.

This chapter provides a general overview of the state-of-the-art in Arabic text recognition technology. Section 2 presents the characteristics of Arabic text with respect to OCR. Section 3 introduces a typical general model for Arabic OCR. Related research on the pre-processing of text images is discussed in Section 4. Section 5 addresses the literature on segmentation of Arabic Text. Common feature extraction techniques are presented in Section 6. Section 7 discusses the used classification methods in Arabic text recognition. The post-processing related research work is addressed in Section 8. Section 9 discusses the available databases for Arabic OCR research. Section 10 lists available Commercial Arabic OCR Software. Section 11 discuses the challenges related to Arabic OCR and discusses possible future trends.

CHARACTERISTICS OF ARABIC TEXT

Arabic is a cursive language written from right to left. It has 28 basic alphabets. An Arabic letter might have up to four different shapes depending on the position of the letter in the word: whether it is a standalone letter, connected only from right (initial form), connected only from left (terminal form), or connected from both sides (medial form). Letters of a word may overlap vertically (even without touching).

Arabic letters do not have fixed size (height and width). Letters in a word can have diacritics (short vowels) such as *Fat-hah, Dhammah, Shaddah, sukoon* and *Kasrah*. Moreover, *Tanween* may be formed by having double *Fat-hah*, double *Dhammah*, or double *Kasrah.* Figure 1 lists these diacritics. These diacritics are written as strokes, placed either on top of, or below, the letters. A different diacritic on a letter may change the meaning of a word. Readers of Arabic are used to

Figure 1. Arabic short vowels (diacritics)