State of the Art in Arabic OCR: Qatari Research Efforts

Mada Center

Introduction

Since the mid-1940s, there has been extensive research and publications on character recognition. With most of the published work being on Latin characters, and Japanese and Chinese characters emerging in the mid-1960s. Despite almost a billion people worldwide using Arabic characters for writing (Arabic, Persian, and Urdu), Arabic character recognition research, starting in the 1970s, is sparse.

This may be attributed to:

- Inadequate journals, books, conferences, funding, and interaction between researchers.
- The lack of utilities like Arabic text databases, dictionaries, programming tools, and supporting staff.
- Delayed onset of Arabic text recognition.
- The techniques developed for other writings cannot be successfully applied to Arabic writing due to the unique attributes of Arabic script.

Arabic OCR challenges

Arabic is written from right to left, which presents many challenges to the OCR developer, which include (Al-Badr 1995; Attia 2004):

The connectivity challenges

Arabic text can only be scripted cursively, i.e., graphemes are connected and only interrupted at limited characters or at the end of the word. This necessitates that any Arabic OCR system undergoes a traditional grapheme recognition task and a more rigorous grapheme segmentation (see Figure 1). To complicate things, both tasks are mutually dependent; therefore, must be done simultaneously.

Figure (1): Grapheme segmentation process illustrated by manually inserting vertical lines at the appropriate grapheme connection points

The dotting challenge

Dotting is extensively used to differentiate characters sharing similar graphemes. Figure (2) shows small differences between members of the same set. Whether the dots are eliminated before the recognition process, or recognition features are extracted from the dotted script, dotting is an area of confusion – therefore, recognition errors – in Arabic font-written OCR systems, especially when using devices such as photocopiers.

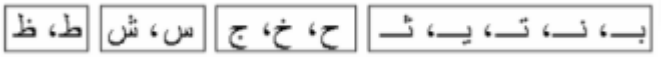

Figure (2): Examples sets of dotting-differentiated graphemes

- The multiple grapheme cases challenge

Due to the connectivity in Arabic orthography, the same grapheme representing the same character can have multiple variants according to its position within the Arabic word segment (Starting, Middle, Ending, Separate) as exemplified by the four variants of the Arabic character "ع" shown in bold in Figure (3).

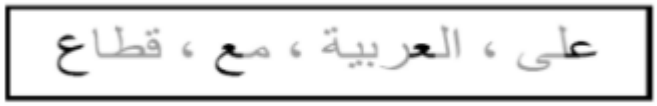

Figure (3): Grapheme "ع" in its 4 positions; Starting, Middle, Ending & Separate

The ligatures challenge

Certain compounds of characters at particular positions of word segments are represented by single atomic graphemes called ligatures; found to some extent in most Arabic fonts. Traditional Arabic font contains around 220 graphemes and Simplified Arabic contains around 151 graphemes. Compared to English with 40 or 50 graphemes. A broader grapheme set means higher ambiguity for the same recognition methodology; hence, more confusion. Figure (4) illustrates some ligatures in Traditional Arabic.

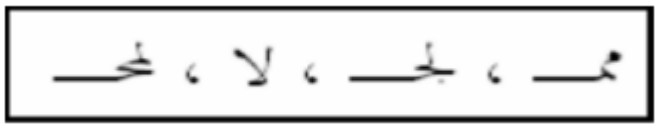

Figure (4): Some ligatures in the Traditional Arabic font

The overlapping challenge

Characters in a word may overlap vertically even without touching, as shown in Figure (5).

Figure (5): Some overlapped Characters in Demashq Arabic font

Size variation challenge

Different Arabic graphemes do not have a fixed height or a fixed width. Moreover, neither the different nominal sizes of the same font scale linearly with their actual line height nor the different fonts with the same nominal size have a fixed line height.

- The diacritics challenge

Arabic diacritics are used only when they help resolve linguistic ambiguity of the text. The problem of diacritics with font written Arabic OCR is that their direction of flow is vertical while the main writing direction of the body Arabic text is horizontal from right to left. (See Figure (6)) Similar dots and diacritics are a source of confusion of font written OCR systems; due to their relatively larger size they are usually pre-processed.

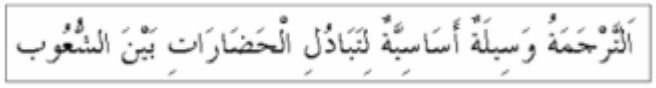

Figure (6): Arabic text with diacritics

Qatari Research Efforts

Arabic Language Technologies Group at Qatar Computing Research Institute (QCRI) is leading research on OCR in Qatar. They are dedicated to promoting the Arabic language by conducting world-class research in Arabic language technologies. Ensuring that the Arabic language flourishes in the digital world is a primary focus area. Some current research projects address the challenges related to the lack of content and extracting that content.
QCRI strives to become the regional and global leader in Arabic language technologies – in the areas of search, information retrieval and analysis, multilingual language processing, advanced machine translation, and leading efforts to increase and enrich Arabic language content online.

Moreover, QCRI's initiatives also examine challenges in retrieving content, making it accessible, and enabling information flow across language barriers. In this regard, development is underway to process Arabic in the search domain such as the use of morphological word analysis, named entity recognition, and data learning technology to detect relevant content for more elaborate analysis. In addition, developing proofing

tools such as typographical checks and language identification for local Arabic dialects and Arabic written using Latin characters.

A major effort at QCRI goes into improving machine translation. Combining an Arabic "Speech-to-Text" engine that permits instantaneous transcription of videos with a machine translation system allows access to broadcast news and news distributed over the web. Future research will concentrate on applications such as lecture translation.

QCRI has established projects related to e-education and making non-native language material accessible. The development of an Arabic language supported e-reader and assistive language tutor are examples that will directly impact society and learning.

Some of the projects run by the Arabic Language Technologies Group at QCRI include:

QATIP – An Optical Character Recognition System for Arabic Heritage Collections in Libraries

The Qatar Computing Research Institute team worked on an end-user oriented QATIP system for OCR in such documents. The recognition is based on the Kaldi toolkit and sophisticated text image normalization. The QATIP interface for libraries consists of a graphical user interface for adding and monitoring jobs and a web API for automated access. It also uses a novel approach for language modeling and ligature modeling for continuous Arabic OCR. The QATIP system was tested on an early print and a historical manuscript and report substantial improvements – e.g., 12.6% character error rate with QATIP compared to 51.8% with the best OCR product (Stahlberg 2015, 2016).

PrepOCRessor

The QCRI Preprocessing Tool for Arabic OCR was developed for preprocessing document images for optical character recognition. A set of image processing operations is chained such that the output of each operation serves as an input to the next one. The tool supports batch processing for high parallelism and scalability. PrepOCRessor is intended to be used in combination with the recognition toolkit Kaldi and supports file formats for feature sets (.ark,t) and forced-alignments (.al) for seamless integration. Though the focus is on Arabic script, the tool has been successfully used for other writing systems, e.g., Latin in the ICDAR2015 Competition HTRtS on historic documents.

References

Stahlberg, F., & Vogel, S. (2015, September). The qcri recognition system for handwritten arabic. In *International Conference on Image Analysis and Processing* (pp. 276-286). Springer, Cham.

Stahlberg, F., & Vogel, S. (2016, April). QATIP--An Optical Character Recognition System for Arabic Heritage Collections in Libraries. In 2016 12th IAPR Workshop on Document Analysis Systems (DAS) (pp. 168-173). IEEE.

Al-Badr, B., & Mahmoud, S. A. (1995). Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), 49-77.

