Overview of Arabic OCR and Related Applications

Mada Center

Optical Character Recognition (OCR) is a generic term used to characterize technologies that recognize text within scanned documents, and photos, to help convert them into a digital format. OCR technology is used to convert virtually any kind of images containing written text (typed, handwritten, or printed) into machine-readable text data. OCR has become a key area of interest over the past two decades with respect to implementing projects related to digitize historic documents (e.g., newspapers, manuscripts, constitutional bills, letters, etc.). The importance of OCR technologies has become even more widespread with the advent of the internet which serves as a resource of multilingual information based on digital textual data.

While OCR technology has undergone several improvements over time and achieved close to a hundred percent accuracy in languages based on Latin scripts (e.g., English), there have been major challenges in enhancing OCR accuracy for languages based on right-to-left reading stylized scripts (e.g., Arabic, Persian, Urdu, etc.). Arabic is the first language for more than 400 million people worldwide and Arabic speaking readers represents a major proportion of internet users who are potentially interested in accessing Arabic digital resource. Hence, the importance of optimizing Arabic OCR technology is significantly vital for improved information and knowledge sharing within society.

The fundamental challenges involving Arabic OCR is the fact that recognition accuracy is harder to achieve primarily due to the following characteristics of the Arabic script set:

- **Character Position:** An Arabic character may have one to four unique shapes depending on its position within a word (i.e., isolated, initial, middle, and end). The OCR solution must be able to effectively identify the concerned Arabic character irrespective of its position within a word.
- **Dot and No-Dot Character:** Certain Arabic characters may have the presence of dots above or under them which can impact the outcome of the final character or word. There may be one to three dots used with the character to determine the final word.
- **Dot Character Baseline:** The presence of a dot within a character is in relation to a baseline as the dot used with the Arabic character may be located above or below the baseline (where applicable). The baseline is significant in developing Arabic OCR systems as it helps to classify Arabic characters into two classes: dot character above the baseline and dot character below the baseline.
- **Zigzag-Shaped Character:** Another distinguishing characteristic of Arabic script is the presence of Hamza, a zigzag-shaped mark (ء) with some Arabic characters which can pose challenges for the OCR systems to recognize the character or word accurately.
- **Loop-Shaped Character:** Several Arabic characters have a loop shape, such as Saad (ص), Dhad (ض), Fa ((ف, Meem (م), and Qaf (ق). An obstacle

for Arabic OCR is to be able to accurately recognize Arabic characters that contain a loop shape.

- **Diacritics:** Some Arabic text may be written with diacritical marks accompanying each character which makes it challenging for the OCR engine to identify the character effectively because this affects the graphical analysis of the image.

Over the past decades, researchers and scientists have worked on developing various databases of Arabic handwritten words to serve as a reference for OCR developers to create solutions for identifying textual shapes, characters, and reconciliating them into a digital text format. In 2002, a database on Arabic handwritten words (IFN/ENIT-database) was made available to the community. In September 2006, a summit on Arabic and Chinese Handwriting Recognition was held at College Park, MD in the US where experts from both research fields presented their actual work. From that time, intensive research on Arabic script recognition started and has resulted in a big step forward today.

The most common application of OCR technology is converting printed paper documents to machine-readable digital text format. Some other areas of application for OCR technology are as follows (but not limited to):

- Data Entry Automation
- Indexing Documents for Search Engines
- Automatic Vehicle Plate Number Recognition
- Voucher Code Scanning
- Office filing system
- Self-service stores / e-Kiosk
- Digitizing handwritten documents, books, and manuscripts
- Assistive Technology

OCR technology has a key role to play in the development of Assistive Technologies to help improve the lives of People with Disabilities (PWDs). As such PWDs, primarily Visually Impaired individuals, cannot use their Assistive Technology to read digital content without the accurate utilization of OCR technology. With improved Arabic OCR, PWDs can enjoy greater access to digital documents and improve their quality of life across education, employment, and other aspects of daily living. Additionally, the availability of digital text is vital to make printed information accessible to PWDs because this enables the creation of information in other accessible formats such as audio, large print, and Braille. Digital text is especially helpful for struggling readers, including those who have learning difficulties such as dyslexia.

Mada Center has a significant role to propel the improvement of Arabic OCR and the development of innovative accessible OCR based solutions. This is done by supporting relevant innovators and entrepreneurs through the Mada Innovation Program to successfully develop their Assistive Technology solutions and prepare them to be market-ready for Qatar and the Arabic speaking region. Mada Center strives to increase the number of ICT Digital Accessibility solutions to adequately serve the growing needs of PWDs in Qatar and the region.

