# Arabic Optical Character Recognition (OCR) Technology at Qatar National Library

Hany A. Elsawy Abdellatif

Optical Character Recognition (OCR) is the practice of extracting text from images. The process itself is becoming popular in terms of usage and research, as it spans multiple areas of science, including image processing, machine learning, information retrieval, and artificial intelligence.

In layman's terms, it is the only way to copy, use, and index the text from a scanned image. The benefits vary from a simple copy/paste process, citation, search within, text annotation, and tagging. Moreover, it perfectly meshes with the modern algorithms of text mining, morphological search, text automatic translation, text summarization, linked data, and indexing tools.

Using OCR images is the ultimate value-add to scanned documents—it brings the documents to life and allows users to discover every bit of information stored within.

At Qatar National Library, multiple techniques and algorithms to OCR text have been developed; these methods include both human operators and automatic SDKs (Software Development Kits) and APIs (Application Programming Interfaces). Adittionally, QNL built an accurate yet scalable system that will efficiently streamline the operation, as it harmonizes the roles and responsibilities between humans and machines to reach the maximum quality of the extracted text.

While we use OCR for a wide array of languages, we are most proud of our achievements with regards to the Arabic text. Since the start of machine learning algorithms, and even with the most modern OCR executables, Arabic text remains a formidable challenge. The multiple shapes and sizes of the Arabic font, in addition to its diacritics, dot usage, cursive characters, and the changing of character shapes based on their location inside a word were all factors that reduced the quality of the OCRed text.

Nafath © 2021 by Mada Center is licensed under CC BY-NC-ND 4.0

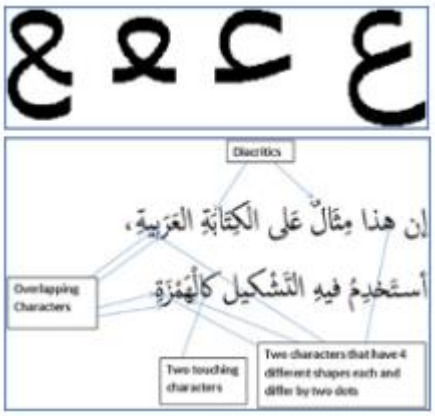


FIGURE 01 Arabic OCR challenges

With the proper tools and algorithms, QNL built universal libraries that cover 99% of the printed Arabic text based on shape, quality, and size and we have engineered a clear workflow to enhance the quality of the images at the start by raising the DPI, smoothing the edges, refining smudged printing and removing the noise. With this image-enhancement process and the trained machine learning libraries, our OCR accuracy achieves 99 percent character level accuracy in Arabic.


FIGURE 02 shape classification

This allowed QNL to index the output text using robust yet sophisticated Arabic text lexical analyzers and offered that to our patrons with just a single click. Our patrons need only to access the Library's Digital Repository and enjoy the "search within" feature, which is expected to immensely help improve the quality of research in Arabic studies, such as art, history, science, and philosophy, to name just a few.

Nafath © 2021 by Mada Center is licensed under CC BY-NC-ND 4.0

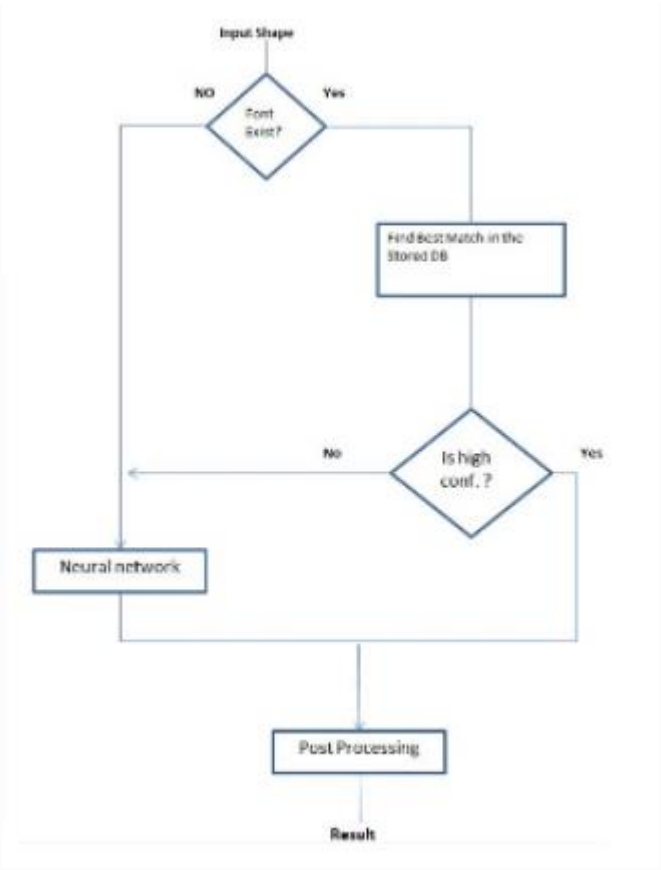


FIGURE 03 Arabic Text Recognition flow

The Library's state-of-the-art digitization facility makes Arabic content from its Heritage Library and other institutions available online, increasing the availability of Arabic content worldwide.

The Library harnesses the expertise of a fully trained international team, and laboratories equipped with cutting-edge technology, to undertake various processes of digital preservation. The Library offers the services of bulk digitization, large-format scanning and image stitching, on-site digitization, E-Pub creation, Optical Character Recognition (OCR), 3D Photography, audiovisual digitization, and long-term preservation.

In addition to ongoing efforts to digitize the Library's collections of rare books, manuscripts, maps, and photographs, the Library's Digitization Center is working on digitization and Arabic OCR projects with other heritage collections in Qatar and international institutions, including:

- New York University (NYU) project: This joint project applies optical character recognition to more than 8,000 Arabic books in NYU library collections, which will also be available on Qatar National Library's online platforms.

Nafath © 2021 by Mada Center is licensed under CC BY-NC-ND 4.0 

- Doha Historical Dictionary of Arabic Language: The Library contributes to the field of optical recognition of Arabic characters, which will assist research on the etymology and meaning of Arabic words.
- Museum of Islamic Art: A Memorandum of Understanding outlined possible collaborations, including a project to digitize **164** of the rarest books and manuscripts in the museum as well as its and library collections, including the Latin OCR.
- Al Shaqab Horse Collection: the Library digitized over 50,000 images from Al Shaqab's Horse Collection.
- Ottoman Archive: 1,600 digital images of heritage documents related to the Gulf region from the Ottoman Archive have been processed, to be made available on the Library's online platforms.
- Qatar Traditional Architecture Photographic Collection: The Library digitized a collection of 1793 photographs from a 1985 French archeological expedition to Qatar that produced a comprehensive record of traditional 19th-century architecture.

The Digitization Center follows international best practices and guidelines, including the Federal Agencies Digitization Guidelines Initiative (FADGI), Metamorfoze Preservation Imaging Guidelines, ISO- 19264, and the IFLA guidelines for digitization projects. This enabled the Library to digitize **10,277,367** pages from various collections including **4,957,546** Arabic pages from Qatar National Library's Heritage Collection, and **2,782,016** pages from the Online Arabic Collection of New York University.

Libraries play an important role in preserving heritage for future generations, and digitization and its sophisticated processes and operations go a long way in ensuring this is done. Furthermore, Qatar National Library is committed to the preservation of heritage not only of the region but of the Islamic world as a whole. We have come a long way in building a reliable process for digitization and OCR Arabic content for the benefit of spreading rich Arabic knowledge and heritage; we are committed to working harder to fulfill that goal.

Nafath © 2021 by Mada Center is licensed under CC BY-NC-ND 4.0