Improving Optical Character Recognition for Historic Arabic Documents

Project Description:

The dense and rich history of the Arab world has been captured over the centuries in various text manuscripts. In this modern digital age, it is important to catalog and present this information digitally such that it can easily be indexed and searched. We develop state-of-the-art tools here at QCRI to perform optical character recognition (OCR) on these historic documents, i.e. write algorithms to convert the scanned images to digital text. A big part of this process is to first identify the layout of a given scanned page in a document, and then find where the text is present in the page while ignoring borders, illustrations and designs that appear along with the text. This project will focus on this specific task of page layout analysis. As an intern, you will research the existing solutions out there for this task, evaluate them and adapt the best one to the specific domain of history Arabic documents.

Duties/Activities:

- Research existing literature and software used for page layout analysis for common languages
- Research existing literature and software used for page layout analysis for historic documents
- Evaluate available options on our existing historic dataset
- Adapt and incorporate the best solution into our existing state-of-the-art pipeline

Required Skills:

- Good Programming Skills
- Comfortable with UNIX command line tools
- Preferred: Knowledge of various Image formats

Learning Opportunities:

- Learn about state-of-the-art techniques in page layout analysis, which is not only used in OCR on text documents, but also in augmented reality translation apps today
- Get hands on experience with a complete OCR pipeline
- Learn research methodologies and the art of assessing and evaluating existing solutions
- Make a meaningful contribution to a tool that is used frequently by libraries to digitize their existing literature

Expected Team Size:

One to two students

Mentors

Name: Fahim Dalvi **Email:** faimaduddin@qf.org.qa

Name: Stephan Vogel **Email:** svogel@hbku.edu.qa