OCRopus++: A High performance OCR System For Medieval Documents

Digitizing historical documents is crucial in preserving the literary heritage. With the availability of low cost capturing devices, libraries and institutes all over the world have old literature preserved in the form of scanned documents. However, searching through these scanned images is still a tedious job as one is unable to search through the written text itself. Current OCR systems, both open-source and commercial ones, have been applied successfully to recognize text in both printed and handwritten form. However, for processing medieval historical documents, complex layout analysis and requirement of a lot of transcribed data for training OCR model are major performance limiting factors for these systems. OCRopus++ is an initiative to solve these problems to a great extent and produce a high performance OCR system for medieval documents.

Syed Saqib Bukhari

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH

Kaiserslautern