Urdu Nastalique Optical Character Recognition

Qurat ul Ain Akram
Sr. Research Associate, Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore

Abstract:

This workshop presents the framework for the development of Urdu Nastalique Optical Character Recognition (OCR) system using Tesseract[1]. Tesseract provides multilingual framework for recognition of text of different scripts. Urdu is cursive in nature and is written using Arabic script. It has bidirectional writing style. In Nastalique, Urdu is written diagonally and has complex marks and dots placement rules. It has character and ligature overlapping. Limited research has been carried out for the recognition of Urdu Nastalique text. This workshop is aimed at presenting Tesseract-based recognition framework for Urdu Nastalique text images. This workshop focuses on giving hands on experience of Tesseract training and recognition of Urdu document images.

[1] http://code.google.com/p/tesseract-ocr/