CLE SEMINAR SERIES-III

Topic: Maturation Process of the Ligature Based Urdu Noori Nastalique Optical Character Recognizer.

Presenter: Aneeta Niazi

Presentation Date: 22nd April, 2014

Venue: KICS Seminar Hall

Abstract:

The ligature based classification and recognition process requires the training of main body images of Urdu ligatures with Tesseract. The time consumed by the manual training process of 5586 main bodies has been considerably reduced by automatically generating the training files. The recognition accuracy of the OCR has been improved by dividing the trained data in to 4 different sets of overlapping recognizers, instead of using a single recognizer. The trained data for 22 and 36 font size recognizers have also been matured on scaled data of 18-28 and 30-44 font sizes respectively. Additional main bodies with attached diacritics have been added in the training data. The Lookup Table generation process has been carried out automatically. This automatically generated lookup table has been incrementally modified by testing and manual analysis of OCR system on document images scanned from Urdu books. The system with matured lookup table has been tested on 199 images, and testing results have shown accuracy improvement from 77% to 87% of classification and recognition module.