CLE SEMINAR SERIES-III

Topic: Font Size Independent Urdu Nastalique Noori Recognition and Classification Using Tesseract: Challenges Faced with Large Font Sizes

Presenter: Ms. Amna Ejaz

Presentation Date: 18th February, 2014

Venue: KICS Seminar Hall

Abstract:

Urdu OCR (Optical Character Recognition) is divided into three stages: pre-processing, recognition and post-processing. The ligature based classification and recognition process requires extraction of the main body images of Urdu ligatures and training those images with Tesseract. Tesseract is an open source OCR engine and has been used in OCRs for many languages. The main body images are prepared through printing ligatures, scanning the printed pages, and extraction of main body images. Once prepared, the data is trained and testing using Tesseract, which extracts features of the main bodies. Large font sizes (28, 30, 32, 34, 36, 38, 40, 42 and 44 font sizes) are scaled to a single font size and recognized. While testing data of large font sizes, different problems are faced as the slightest variation in shape is noticeable in images.