CLE SEMINAR SERIES-III

Topic: Page Segmentation in Urdu Nastalique Optical Character Recognizer

Presenter: Mr. Qasim Ali

Presentation Date: 29th September, 2014

Venue: KICS Seminar Hall

Abstract:

To process a document image in Optical Character Recognizer (OCR), the scanned image is read and broken down into logical entities such as text area, line, word and character. This process is known as page segmentation. Various approaches have been employed to segment a page;top down approach is commonly used where histograms are used to find segmentation points based on peaks and valleys to break a page into lines and further into connected components or characters. Another approach is the bottom up where connected components are first extracted and then using the positional information lines and columns are formed. Cursive nature of Nastalique poses various challenges in segmentation and in the project[1], two-level segmentation approach has been devised where first a layout of page is found which marks the text areas and non-text areas. Later bottom up approach is implemented to segment text area into lines.

[1] www.urduocr.net