

Localization, Extraction and Recognition of Text in Telugu Document Images

Atul Negi
Department of CIS
University of Hyderabad
Hyderabad 500046, India
atulcs@uohyd.ernet.in

K. Nikhil Shanker
Department of CSE
Mahatma Gandhi
Institute of Technology
Hyderabad, India
nikhil.shanker@acm.org

Chandra Kanth Chereddi
Department of CSE
College of Engineering
Osmania University
Hyderabad 500007, India
chandra-kanth@ieee.org

Abstract

In this paper we present a system to locate, extract and recognize Telugu text. The circular nature of Telugu script is exploited for segmenting text regions using the Hough Transform. First, the Hough Transform for circles is performed on the Sobel gradient magnitude of the image to locate text. The located circles are filled to yield text regions, followed by Recursive XY Cuts to segment the regions into paragraphs, lines and word regions. A region merging process with a bottom-up approach envelopes individual words. Local binarization of the word MBRs yields connected components containing glyphs for recognition. The recognition process first identifies candidate characters by a zoning technique and then constructs structural feature vectors by cavity analysis. Finally, if required, crossing count based non-linear normalization and scaling is performed before template matching. The segmentation process succeeds in extracting text from images with complex Non-Manhattan layouts. The recognition process gave a character recognition accuracy of 97%-98%.