Abstract

Preprocessing is an important step in the development of Optical Character Recognition (OCR) system. In preprocessing there are various modules like binarization, skew detection and correction etc. Among these modules this paper discusses about binarization module. Although there are many algorithms for binarization of a document image, there are fewer algorithms for binarization of printed color images because of printed color documents contain dither patterns, normal text, reversed text, colored text overlayed on colored background drawings and graphics appear with millions of different colors. Hence preprocessing for colored documents is a challenging task to work. For printed color documents, elimination of dither patterns using Butterworth band reject filter and text extraction in the color documents by eliminating graphics using height of the component is also presented. Results on a corpus consisting of newspapers published in Telugu show that the proposed method shows promising results.
7. Chun-Ming Tsai Intelligent region-based thresholding for color document images with highlighted regions Pattern Recognition 45 (2012) pages 1341-1362

Index Terms

Computer Science        Pattern Recognition

Keywords

OCR, threshold, binarization, dither patterns, connected components