Abstract

The objective of this paper is to propose a new methodology for text extraction and non text removal from colored images. Text in web pages, library documents etc. is one powerful source of high level semantics. Existing text extraction methods do not work efficiently in case of images with complex background, different contrast or text embedded in a complex background. Documents fed into OCR work efficiently if they contain only text. The paper has taken into focus several images in several languages (English, Telgu, and Gurumukhi). Several existing text detection techniques have also been discussed in the paper. The approach used is based on preprocessing steps, adaptive thresholding, detecting connected components, generating blobs and finally extraction of only those blobs which consist of textual part.

References


  - Efthimios Badekas, Nikos Nikolaou, Nikos Papamarkos "Font and Background Color Independent Text Binarization", 2007 Wiley Periodicals, Inc
  - Keechul Jung, Kwang In Kim, Anil K. Jain "Text Information Extraction in Images and Video: A Survey

Index Terms

Computer Science

Keywords

Binarization Pixel Image Text Non Text Text Localization Connected Component Blobs Color