Abstract

In Document Image analysis the digitized images of printed documents typically consist of a mixture of text, graphics, and image elements. For proper processing and efficient representation, these elements have to be separated. For most of the applications it is essential to separate between text and non-text, because text captures the most information. These text lines may have different orientations or the text lines may be of curved shapes. Some of the techniques proposed for text string extraction are completely independent from text orientation and may deal with text in various font styles and sizes. There are many fast and efficient methods for extracting graphics and text paragraphs from printed document. This paper outlines the comparisons of some text extraction techniques proposed by researchers.

References

- Frank Hones, Jiirgen Lichter. "TEXT STRING EXTRACTION WITHIN MIXED-MODE DOCUMENTS" 1993 IEEE.
- Xuhong Li, Peter A. Ng. "A DOCUMENT CLASSIFICATION AND EXTRACTION SYSTEM WITH LEARNING ABILITY"; 2001 IEEE.
- T. Perroud, K. Sobottka, and H. Bunke. "Text extraction from color documents - clustering approaches in three and four dimensions"; 2001 IEEE.
- Jiangying Zhou, Daniel Lopresti. "EXTRACTING TEXT FROM WWW IMAGES"; 1997 IEEE.
- Xuewen Wang. "CHARACTER EXTRACTION AND RECOGNITIONS IN NATURAL SCENE IMAGES"; 2001 IEEE.

**Index Terms**

Computer Science  
Emerging Trends in Technology

**Keywords**

Directed Weight Graph Technique  
Mixed Mode Technique  
Histogram-based Clustering Technique  
Clustering