Abstract

Line and word segmentation is one of the important step of OCR systems. In this paper we have proposed a robust method for segmentation of individual text lines based on the modified histogram obtained from run length based smearing. A complete line and word segmentation system for some popular Indian printed languages is presented here. Both foreground and background information are used here for accurate line segmentation. There may be some touching or overlapping characters between two consecutive text lines and most of the line segmentation errors are generated due to touching and overlapping character
occurrences. Sometimes, interline space and noises make line segmentation a difficult task. Our method can take care of this situation accurately. Word segmentation from individual lines is also discussed here. We have tested our method on documents of Bangla, Devnagari, Kannada, Telugu scripts as well as some multi-script documents and we have obtained encouraging results from our proposed technique.

Reference


2004.

Index Terms

Computer Science  Pattern Recognition

Key words

Word segmentation  Line segmentation

Histogram

Indian documents