Abstract

Text-line segmentation is an essential part of script identification technique from handwritten and printed document images. In case of handwritten documents, overlapping, touching, skewed and small perforations between the lines makes line extraction difficult task. Presence of such variations leads to errors and wrong identification of script. This paper describes an efficient line extraction technique from handwritten document images using histogram and connected component analysis. Using horizontal histogram profile, a threshold, i.e., average
height of a line in the given document is computed, using which the non-overlapping lines are extracted. In order to extract overlapping lines, that exceed the given threshold, a rectangular bounding box is imposed over the words of the overlapping lines using connected component analysis. The mid-point of each bounding box is then calculated and compared with the average height of the image to label each component as either belonging to upper line or lower line. Experiments are carried out on document images of Kannada, Telugu, Hindi, English and Malayalam scripts and the results obtained are encouraging.

References

Index Terms

Computer Science

Pattern Recognition
Keywords
Handwritten Document  Text-line  Segmentation  Histogram  Connected Component.