Abstract

Based on the analysis of Indian script character shapes and literature survey, it presents a new sequence of line and word segmentation method to handle some of the deformations usually present in the handwritten document like touching components, overlapping components, skewed lines, words with individual skews etc. and build a proper text image with all these deformations removed. Line segmentation procedure is applied using Hough transform. The
word segmentation is done with the computation of the distances of adjacent components in the text line image and classification of the previously computed distances as either inter-word gaps or inter-character gaps in a Gaussian mixture modeling framework. The proposed method of line segmentation is a sufficiently accurate to extract the text lines from unconstrained handwritten text documents. Word segmentation procedure also works well on different language scripts. Average result of word segmentation for complex Document on different language script is 76% and average result of word segmentation for good Document of different language script is 90%.

References


Index Terms

Computer Science
Emerging Trends in Technology
Keywords
Optical character recognition  Pre-processing  Global skew detection and correction
Line segmentation

Word segmentation