Abstract

Text and not-text segmentation and text line extraction from document images are the most challenging problems of information indexing of Arabic document images such as books, technical articles, business letters and faxes in order to successfully process them in systems such as OCR. Researches on Arabic language related to documents digitization have been focusing on word and handwriting recognition. Few approaches have been proposed for layout analysis for Arabic scanned/captured documents. In this paper we present a page segmentation method that deals with the complexity of the Arabic language characteristics and fonts using the combination between two algorithms. The first method is the Run length Smoothing. The second method is the Connected Component Labeling algorithm for text and non-text classification using SVM. The combination of the two methods is based on Anding and Oring operations between the outputs of the two methods based on certain conditions. Then, dynamic horizontal projection based on dynamic updating of the threshold to commensurate with the noise associated with different documents and in between text lines. The performance evaluation is performed using manually generated ground truth representations from a dataset of Arabic document images captured using cameras and a hardware built for this purpose.
A Combined Algorithm for Layout Analysis of Arabic Document Images and Text Lines Extraction

Evaluation and experimental results demonstrate that the proposed text extraction method is independent from different document size, text size, font, shape, and is robust to Arabic document segmentation and text lines extraction.

References


Index Terms
Computer Science
Pattern Recognition

Keywords
Arabic document layout analysis
Arabic text line extraction