International Journal of Electronics and Communication Engineering
Vol:8, No:11, 2014

Persian/Arabic document Segmentation Based on Pyramidal Image Structure

Authors : Seyyed Yasser Hashemi, Khalil Monfaredi

Abstract : Automatic transformation of paper documents into electronic documents requires document segmentation at the first stage. However, some parameters restrictions such as variations in character font sizes, different text line spacing, and also not uniform document layout structures altogether have made it difficult to design a general-purpose document layout analysis algorithm for many years. Thus in most previously reported methods it is inevitable to include these parameters. This problem becomes excessively acute and severe, especially in Persian/Arabic documents. Since the Persian/Arabic scripts differ considerably from the English scripts, most of the proposed methods for the English scripts do not render good results for the Persian scripts. In this paper, we present a novel parameter-free method for segmenting the Persian/Arabic document images which also works well for English scripts. This method segments the document image into maximal homogeneous regions and identifies them as texts and non-texts based on a pyramidal image structure. In other words the proposed method is capable of document segmentation without considering the character font sizes, text line spacing, and document layout structures. This algorithm is examined for 150 Arabic/Persian and English documents and document segmentation process are done successfully for 96 percent of documents.

Keywords : Persian/Arabic document, document segmentation, Pyramidal Image Structure, skew detection and correction
Conference Title : ICEP 2014 : International Conference on Electronic Publications
Conference Location : journal city, WASET
Conference Dates : November 23-23, 2014