عنوان مقاله:

Document Analysis And Classification Based On Passing Window

نویسنده:

ZAHER BAMASOOD - *Computer Science Department, Hadhramout University*

خلاصه مقاله:

In this paper we present Document analysis and classification system to segment and classify contents of Arabic document images. This system includes preprocessing, document segmentation, feature extraction and document classification. A document image is enhanced in the preprocessing by removing noise, binarization, and detecting and correcting image skew. In document segmentation, an algorithm is proposed to segment a document image into homogenous regions. In document classification, Neural Network (Multilayer Perceptron- Back propagation) classifier is applied to classify each region to text or non text based on a number of features extracted in feature extraction. These features are collected from different other researchers' works. Experiments were conducted on 398 document images selected randomly from printed Arabic text database (PATDB) which was selected from various printing forms which are advertisements, book chapters, magazines, newspapers, letters and reports documents. As results, the proposed segmentation algorithm achieved only 0.814% as ratio of the overlapping areas of the merged zones to the total size of zones and 1.938% as the ratio of missed areas to total size of zones. The features, that show the best accuracy individually, are Background Vertical Run Length (RL) Mean, and Standard Deviation of foreground.

کلمات کلیدی:

Information Retrieval, Document Image Analysis, segmentation, Feature Extraction, data mining

لینک ثابت مقاله در پایگاه سیویلیکا:

https://civilica.com/doc/1030111