UNIVERSITI TEKNOLOGI MARA

THE DOCTORAL RESEARCH ABSTRACTS

Volume: 11, Issue 11 April 2017

ELEVENTH ISSUE

INSTITUTE of GRADUATE STUDIES

IGS Biannual Publication

20

Name : HAMAM M.IBRAHIM MOKAYED

Title : ROBUST GENERIC STRUCTURED DOCUMENT CLASSIFICATION SYSTEM

Supervisor : PROF. DR. HJH. AZLINAH HJ. MOHAMED **(MS)**

The Structured Document Classification System (SDCS) is an industrial-driven technology that has the ability to classify piles of structured documents collected everyday efficiently in different places. Although the SDCS technology has advanced tremendously, one of the most challenging tasks is to propose a classifier that supports various layouts for different categories and different script languages in a high accuracy and efficient time. To solve the issue of supporting various layouts for different categories and different script languages, a Robust Generic Structured Document Classifier has been proposed (RGSDC). RGSDS starts with finding the best objects that can be used to fit the target and solve the issue. Detailed study for all the previous thresholding techniques is conducted to introduce a new categorization method based on the transformation value of input images. This study is a good base for finding reliable thresholding algorithm. A new thresholding technique based on ordinal structure fuzzy logic (OSFM) is proposed to provide a robust generic image thresholding technique (RGT) that is able to extract clear mixed predefined objects for different languages and multi layouts problems. Two different set of features that distinguish different languages and multi layouts structured documents are proposed. Lines-based features are totally relying on dimensions, locations, and slanting properties. On the other hand, blob-based features represent the shape, appearance, and distribution of the nominated objects. All the collected features are used to build a statistical feature vector for the classification stage. Based upon the need of a fast and accurate process to propose a practical structured document classification, a creative and fast skewing process based on nominated referencing lines out of the group of lines in the object selection process is created as the first stage. An algorithm focused on evaluating performance of different pattern classification techniques namely; neural network, support vector machine, Pearson correlation, and dynamic time wrapping (DTW) is used. At the end, computational calculations to prove the accuracy of the proposed algorithms are performed in four different stages. The results of experiments demonstrate that the proposed RGSDC is capable of performing classification of 3.5 forms per second with a 4.9% misclassification error rate (ME). Reliability of the research is verified by benchmarking the results of RGSDS with another well-known SDCS developed by Austrian banking solution company (xyzmo).