World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:10, 2018

Model-Based Field Extraction from Different Class of Administrative Documents

Authors : Jinen Daghrir, Anis Kricha, Karim Kalti

Abstract : The amount of incoming administrative documents is massive and manually processing these documents is a costly task especially on the timescale. In fact, this problem has led an important amount of research and development in the context of automatically extracting fields from administrative documents, in order to reduce the charges and to increase the citizen satisfaction in administrations. In this matter, we introduce an administrative document understanding system. Given a document in which a user has to select fields that have to be retrieved from a document class, a document model is automatically built. A document model is represented by an attributed relational graph (ARG) where nodes represent fields to extract, and edges represent the relation between them. Both of vertices and edges are attached with some feature vectors. When another document arrives to the system, the layout objects are extracted and an ARG is generated. The fields extraction is translated into a problem of matching two ARGs which relies mainly on the comparison of the spatial relationships between layout objects. Experimental results yield accuracy rates from 75% to 100% tested on eight document classes. Our proposed method has a good performance knowing that the document model is constructed using only one single document.

Keywords : administrative document understanding, logical labelling, logical layout analysis, fields extraction from administrative documents
Conference Title : ICIAP 2018 : International Conference on Image Analysis and Processing
Conference Location : Paris, France
Conference Dates : October 29-30, 2018