World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:11, 2020

Visual Template Detection and Compositional Automatic Regular Expression Generation for Business Invoice Extraction

Authors : Anthony Proschka, Deepak Mishra, Merlyn Ramanan, Zurab Baratashvili

Abstract : Small and medium-sized businesses receive over 160 billion invoices every year. Since these documents exhibit many subtle differences in layout and text, extracting structured fields such as sender name, amount, and VAT rate from them automatically is an open research question. In this paper, existing work in template-based document extraction is extended, and a system is devised that is able to reliably extract all required fields for up to 70% of all documents in the data set, more than any other previously reported method. The approaches are described for 1) detecting through visual features which template a given document belongs to, 2) automatically generating extraction rules for a given new template by composing regular expressions from multiple components, and 3) computing confidence scores that indicate the accuracy of the automatic extractions. The system can generate templates with as little as one training sample and only requires the ground truth field values instead of detailed annotations such as bounding boxes that are hard to obtain. The system is deployed and used inside a commercial accounting software.

Keywords : data mining, information retrieval, business, feature extraction, layout, business data processing, document handling, end-user trained information extraction, document archiving, scanned business documents, automated document processing, F1-measure, commercial accounting software

Conference Title : ICDAR 2020 : International Conference on Document Analysis and Recognition

Conference Location : Jerusalem, Israel

Conference Dates : November 26-27, 2020