smartFIX - The Multichannel Document Analysis Product by Insiders Technologies

Florian Deckert, Michael Gillmann
Insiders Technologies GmbH
Brüsseler Straße 1, 67657 Kaiserslautern, Germany
{*f.deckert, m.gillmann*}@*insiders-technologies.de*

smartFIX [3] is a document analysis product for knowledge-based extraction of data from any document format. Paper documents as well as any type of electronic document format (e.g. faxes, e-mails, MS Office, PDF, HTML, XML, etc.) can be processed. Regardless of document format and structure, smartFIX recognizes the document type and any other important information during processing.

smartFIX imports the documents to be processed from various sources. Scanned paper documents, incoming fax documents, e-mails, and other electronic documents are processed. Basic image processing, like binarization, despeckling, rotation and skew correction is performed on each image page. If desired, smartFIX automatically merges individual pages into documents and creates processes from individual documents. For each document, its document class is determined, defining the business process to be triggered in the company. smartFIX subsequently identifies all relevant information on the document belonging to the respective business process. In this step, smartFIX can use customer relation and enterprise resource planning data (ERP data) provided by a matching database and other sophisticated knowledge-based methods to increase the detection rate. The quality of extracted data is enhanced by automatic mathematical and logical checks over all recognized values. To do this e.g. Constraint Solving [2] and Transfer Learning methods [4] are used. Values that are accurately and unambiguously recognized are released for direct export; uncertainly [5] recognized values are forwarded to a verification workplace for manual checking and verification. The quality-controlled data is then exported to the desired downstream systems e.g., an enterprise resource planning system like SAP for further processing.

smartFIX provides self-learning mechanisms as a highly successful method for increasing recognition rates. The self-learning mechanisms use the post-verification quality-controlled data as ground truth (GT) in order to find rules for the analysis step. Data that has been validated automatically is used to evaluate the reliability of the learned rules as well. Typical learned rules are the position of fields relative to stationary layouts or keywords, regular expressions, relative positions of extracted information (e.g. net amount and total amount) and many more.

smartFIX uses a strategy that searches for all entries contained in the customers database on the document. The procedure is independent of location, layout and completeness of the data on the document. Within smartFIX this strategy is called “Top Down Search”. It supplies results normally within less than one second even on large databases.

Many of the documents processed in smartFIX contain tables of vital information. Examples are position tables on invoices or tables containing requested items on orders. Our experience is that there is no clear layout that helps to identify columns and the BP only needs a subset of the provided information. Therefore smartFIX does not only rely on a physical structure. Table extraction in smartFIX is based on expectations about the presence and semantics of certain data entities in order to understand a table’s content [1].

Acknowledgment

The work presented in this paper was performed in the context of the Software-Cluster project EMERGENT (www.software-cluster.org). It was partially funded by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IC10S01. The authors assume responsibility for the content.

References

[1] F. Deckert, B. Seidler, M. Ebbecke, and M. Gillmann, *Table Content Understanding in smartFIX*, 11th Int. Conf. on Document Analysis and Recognition (ICDAR), Beijing, China, 2011.

[2] A. Fordan, *Constraint Solving over OCR Graphs*, 14th Int. Conf. on Applications of Prolog (INAP), Tokyo, Japan, 2001.

[3] B. Klein, A. Dengel, and A. Fordan, *smartFIX: An Adaptive System for Document Analysis and Understanding*, in: A. Dengel, M. Junker, A. Weissbecker (Eds.), *Reading and Learning - Adaptive Content Recognition*, LNCS 2956, Springer, 2004.

[4] F. Schulz, M. Ebbecke, M. Gillmann, B. Adrian, S. Agne, and A. Dengel, *Seizing the Treasure: Transferring Layout Knowledge in Invoice Analysis*, 10th Int. Conf. on Document Analysis and Recognition (ICDAR), Barcelona, Spain, 2009.

[5] B. Seidler, M. Ebbecke, and M. Gillmann, *smartFIX Statistics – Towards Systematic Document Analysis Performance Evaluation and Optimization*, 9th IAPR Int. Workshop on Document Analysis Systems (DAS), Boston, MA, USA, 2010.