2009 10th International Conference on Document Analysis and Recognition

Seizing the Treasure: Transferring Layout Knowledge in Invoice Analysis

Frederick Schulz, Markus Ebbecke, Michael Gillmann
Insiders Technologies GmbH
[f.schulz,m.ebbecke,m.gillmann]@insiders-technologies.de

Benjamin Adrian, Stefan Agne, Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
[firstname.lastname]@dfki.de

Abstract

This paper deals with the transfer of knowledge on invoice document layout and extraction strategies. This knowledge has been automatically generated by self-teaching mechanisms of the invoice analysis software smartFIX over several years of operation. We present results of analyzing this "treasure" of knowledge and putting it to use in smartFIX systems of new users. The evaluation shows that this transfer of knowledge using state-of-the-art techniques in transfer learning achieves significantly higher initial recognition rates than the unaugmented system, delivering instant economic advantages by reducing accountant personnel workload.

1. Introduction & Motivation

Every company has to deal with invoices for received goods or services. The majority of these is still printed on paper, partly due to legislative impediments on the use of electronic documents for invoicing. Manual invoice processing is both expensive (Klein et al. [10] estimate costs of € 9 per invoice) and error prone. Therefore, automatic analysis of invoices is a key process in rationalizing the accounts payable workflow and reducing the number of accounting personnel occupied with it. Several software companies offer solutions and services in this field. As a result of the large variance in invoice layouts, systems that use manually defined models for each class of invoices have become obsolete. State-of-the-art systems use machine learning technology to automatically derive extraction strategies [10].

The German software company Insiders Technologies GmbH offers an invoice analysis system (IAS), called smartFIX INVOICE [9], which has its origins in research conducted at DFKI. Since its introduction on the market in 2001, over 200 enterprises have started using smartFIX, each processing hundreds to several ten thousands of invoices every day. Some of these invoices are verified by accounting clerks, producing ground truth data and validating the extraction strategy and parameters used. The smartFIX system exploits those data samples extensively in deriving vendor-specific extraction strategies. This learning process accumulates a real "treasure" of knowledge, enabling smartFIX to achieve impressive rates of recognition of more than 98% correctness and virtually no false positives.

Until now, each smartFIX installation creates this knowledge pool on its own, starting with acceptable, but not exceptional rates of recognition and therefore much higher workload for the verifying and post-processing accountants. The recognition rates converge reasonably fast towards the optimum and the verification workload quickly declines, but, nevertheless, this additional effort is pricey and avoidable. This is where the work presented here fits in.

In a joint project with DFKI, Insiders Technologies has researched ways to transfer the knowledge of invoice layouts and extraction strategies to smartFIX systems of other customers. Special notion was given to the strict protection of data privacy of both smartFIX users and their vendors. This emphasis on privacy is necessary because layout and analysis data samples are intimately linked to the customer relation databases of each user.

2. Related Work

Invoice analysis systems (IAS) have been a research topic for years [1, 10, 11]. smartFIX [6] is a state-of-the-art IAS that has been developed by Insiders Technologies GmbH. This work extends smartFIX by using techniques for transfer learning to use earlier acquired knowledge about invoice document layout and extraction strategies in a current application domain.

A comparable approach that uses a multi-class classifier for an IAS based on keywords and layout anchors is presented by Cesarini et al. [5]. The au-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.47

thors distinguish between two knowledge levels: *Class-Independent Domain Knowledge (CIDK)* and *Class-Dependent Domain Knowledge (CDDK)* [4]. A class consists of invoices issued by the same company (vendor). The CIDK describes layout and logical similarities of all invoices of the domain. The CDDK represents the similarities of the invoices belonging to a specific class. Their system uses only CIDK for unclassified invoices.

The work described in this paper extends smartFIX by a CBR technique in order to retrieve the most similar known invoice to an incoming unknown invoice.

The basics of this idea have been presented by Hamza et al. in [7, 8], where CBR techniques work on probing layout graphs by edit distances [3, 12]. The comparison of existing classes with an incoming invoice resulted in a ranked list of invoice classes with similar layouts.

The suitability of CBR methods for document analysis and understanding has also been proven by Beusekom et al. [2]. Here CBR methods are used for reusing labeled documents in a logical labeling task.

With the extension described in this document, a running smartFIX system retrieves existing extraction rules about a class of invoice cases. Given the most similar class of invoices with respect to layout, existing extraction rules are used to extract information even on yet unknown invoice classes.

In the next section, some essential aspects of the smartFIX system are presented that are necessary to understand the following approach.

3. smartFIX System Architecture

smartFIX is a product portfolio 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 page image. If desired, smartFIX automatically merges individual pages into documents and creates processes from individual documents. For each document, the document class and thus the business process to be triggered in the company is implicitly determined. smartFIX subsequently identifies all relevant data contained in the documents and related 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 to increase the detection rate. An automatic quality check is then performed on all recognized values. Both general (CIDK) and vendor-specific (CDDK) rules are used. Values that are accurately and unambiguously recognized are released for direct export; "uncertain" (inaccurately) 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. An overview of the system architecture can be seen in figure 1.

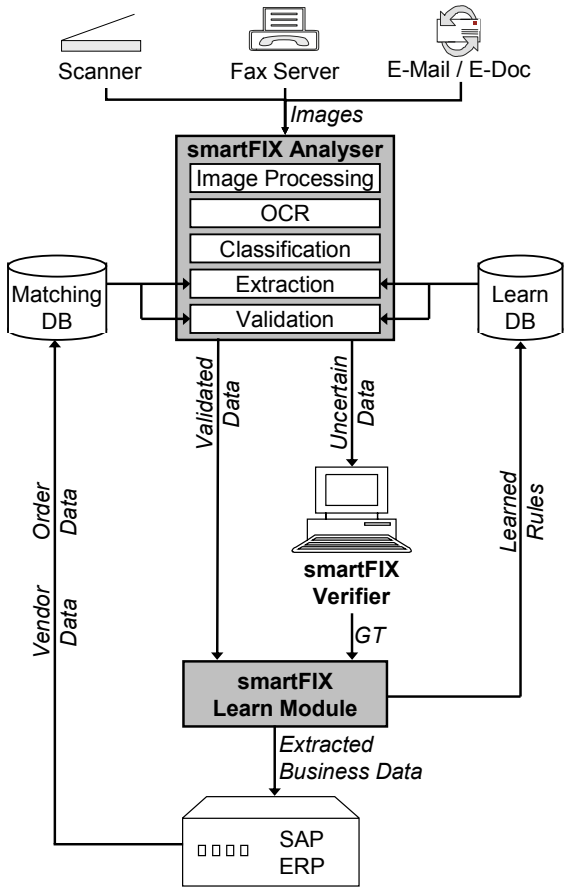

Fig. 1: smartFIX system architecture

smartFIX provides self-teaching mechanisms as a highly successful method for increasing recognition rates. The self-teaching 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. Thus, every invoice processed contributes to the enlargement of the CDDK. 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.

For invoice processing learned rules are linked to the vendor. To select the appropriate learned rules during the analysis, the vendor issuing the invoice at hand is extracted. This concept requires a highly reliable vendor search. A special search strategy searches for all entries from the customer's vendor database on the document. The procedure works independently of the 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.

4. Concept

The smartFIX system benefits from knowledge about previously processed documents to generate rules for more accurate analysis. Therefore, a new system has to run for a while until it achieves highly accurate analysis results. For most invoices arriving in such a new system, similar ones have been processed elsewhere in the past. So, their optimal extraction rules have already been learned and stored in the system of a different customer. Currently, there is no way to access this "treasure" of layout knowledge distributed over smartFIX systems of several customers. Transferring these rules to a new system would greatly increase the initial recognition performance. Unfortunately, that information is closely linked to the vendor database and, thus, impossible to transfer for privacy and secrecy reasons. So a new and anonymous way of transferring knowledge of invoice layouts and extraction rules had to be found.

The problem of anonymously storing, transferring, and identifying invoice layouts and extraction rules is solved by a strategy called **layout graphs**. The known document layouts – the **case base** – and the document to be analyzed – the **candidate document** – are transformed into graph representations. Then, the layout graph most similar to the candidate document is determined. Node labels and attributes encode the extraction strategy in each case base graph. The strategy described below is limited to invoice number and invoice date information, but is extendable to other information.

For both candidate documents and ground truth case base documents, the layout graph is derived as follows:

The document image is segmented into word bounding boxes and OCR is performed on those to acquire the textual content. Then the word boxes are classified according to their textual content, depending on the document language. Using a sophisticated fuzzy matching algorithm, the classification is hardly affected by OCR and separation errors. All boxes containing keywords indicating the nearby presence of date and number information are represented by nodes labeled *numberkey* or *datekey*, respectively. Those nodes are called **key nodes**. All boxes containing number and date entries are represented by nodes labeled *number* and *date*, respectively. Those nodes are called **data nodes.** An edge is created between a key node and a data node or another key node, if the second node is exactly above, below, left, or right of the first node. The edge is labeled *above*, *below*, *left*, or *right,* accordingly.

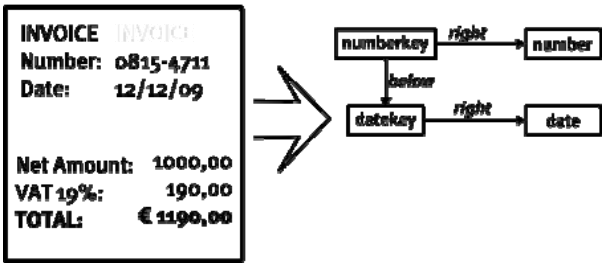

Fig. 2: Conversion from invoice document to layout graph

A natural measure for the similarity of two graphs is their edit distance [3]. Since edit distance computation is very expensive, the **graph probe** approximation [12] is used. The graph probe misestimates the correct edit distance by a factor of four at most. It can be precomputed for all case base documents at setup time and has to be computed only once at runtime: for the candidate document. Two documents can then be compared by a series of simple additions, making a case base of several thousands of documents searchable in negligible time.

The layout graph in the case base having the smallest graph probe distance to the candidate layout graph is retrieved. This graph yields the layout description and extraction rules for the document at hand. Depending on the similarity, a confidence

value is assigned to the output of this step, marking the document for manual or automatic postprocessing.

Extraction rules are described by the constellation of keywords (represented by key nodes) surrounding the desired information (data node). Transferring these rules from the case base layout graph to the candidate document involves mapping those keywords to similar words on the candidate document. Criteria for similarity are string edit distance and deviation from the relative positions in the original constellation. The transferred constellation then reveals the location of the desired data entry. If the keyword constellation transfer fails, the data node position is mapped directly.

Since this learning strategy is mainly used for bridging the gap between installation and optimal performance of the proven and established vendor-based learning strategy described in section 3, there is no need for a case base update mechanism. Instead, a case base created from a large collection of real-world sample invoices is installed once during system setup. This avoids dealing with solution quality assessment and case base maintenance problems.

5. Evaluation and Results

Evaluation was performed on three invoice document corpora: The first and second are collections of 415 (corp1) and 1489 (corp2) images of real-world trade invoices and corresponding ground truth data, acquired from customers using the smartFIX INVOICE product. These corpora represent a cross section of invoices received by major companies in the construction, automotive, and electronics industry. The third corpus (retail) consists of 724 invoices received by a major European retail store chain.

Each corpus contains black-and-white, grayscale, and color document images and the appropriate ground truth data. The images are taken directly from the scan application, no rotation, skew correction, or binarization is applied. For each corpus, the learn data layout graphs were generated and their graph probes computed to fill the case base.

Additional documents from various sources form the case base. It contains a total of 3550 invoice document layout graphs. Care was taken that no test corpus document can be matched to its own layout graph. The construction of the case base took 170 minutes. Thus, the generation of one layout graph takes 3 seconds on average. Those time measurements where taken on a 2.7 GHz processor with 3GB RAM.

Each test image passed through a modified analysis workflow. All extraction mechanisms interfering with the new algorithms were disabled, especially all self-teaching strategies. Instead, the new layout graph strategy with the case base described above was used. The analysis results were recorded and compared to the ground truth data (table row “improved”). Figure 3 shows the results of those test runs.

For comparison, the same images were analyzed on an out-of-the-box smartFIX system without modifications and no vendor-based learn data (table row “unmodified”). This imitates a “fresh” smartFIX installation. These recognition rates are also included in figure 3. Figures 4 & 5 compare the recognition rates for invoice number and invoice date.

The results show a clear improvement: The recognition rates (fig. 4 & 5) are significantly higher after the integration of the layout case base. An even larger case base could be used in real installations, since the analysis times were not much higher with the present case base size.

Direct comparisons to other invoice analysis systems – e.g. those described in [5], [7] and [8] – were not possible, due to the lack of common image and ground truth data collections.

6. Conclusion and Outlook

In this paper, an approach to transfer knowledge about invoice layouts in the form of extraction strategies between users of the invoice analysis system smartFIX was described. The transfer was realized by creating a document layout case base containing the layout descriptions and extraction strategies. Labeled directed graphs were used to identify and compare layouts. As the distance measure for identifying the most similar layout, the edit distance, approximated by graph probing, was employed.

Evaluation with a case base of 3550 invoices on three corpora shows that the knowledge transfer increased the recognition rate by up to 20 percentage points. This led to the decision to include the layout case base strategy in future smartFIX releases.

New smartFIX installations will then contain knowledge about thousands of different invoice layouts from the start. This wealth of knowledge greatly increases the initial performance of those. Even before the vendor-based self-teaching mechanisms reach their full efficiency, most of the information is extracted correctly. This drastically reduces verification workload during the adaption phase and whenever yet unknown or changed invoices arrive. The use of this “treasure” of invoice extraction knowl-

edge may even make the use of smartFIX profitable for smaller enterprises not yet employing an IAS, thus opening new markets. This innovative feature helps Insiders Technologies to stay on the cutting edge of "making documents work".

Further work could be dedicated to compensate scaling, translation, and skewing of the candidate document in the layout tree. These deviations are already determined in the workflow and could further increase the performance on low quality input. An extension of the layout graph to include total and partial amounts, tax rates and amounts or customer-specific information can be realized quickly as the need arises.

corp1 (n = 415)				
	invoice date		**Invoice number**	
	correct	%	Correct	%
unmodified	343	82,7	298	71,8
improved	**402**	**96,9**	**376**	**90,6**

corp2 (n = 1489)				
	invoice date		**Invoice number**	
	correct	%	Correct	%
unmodified	1183	79,5	814	54,7
improved	**1333**	**89,5**	**999**	**67,1**

retail (n = 724)				
	invoice date		**Invoice number**	
	correct	%	Correct	%
unmodified	523	72,2	361	49,7
improved	**638**	**88,1**	**519**	**71,7**

Fig. 3: Evaluation results

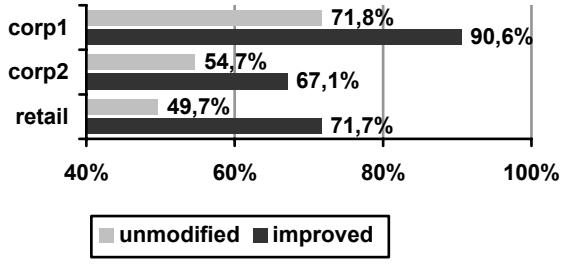

Fig. 4: Recognition rates for invoice number extraction

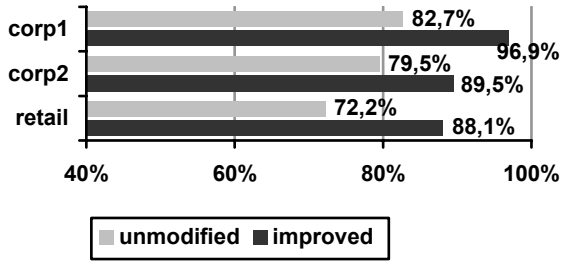

Fig. 5: Recognition rates for invoice date extraction

7. References

[1] T. A. Bayer, H. U. Mogg-Schneider, "A Generic System for Processing Invoices" In: *4th Int. Conf. on Document Analysis and Recognition*, **1997,** p. 740–744

[2] J.v. Beusekom, D. Keysers, F. Shafait, T.M. Breuel, "Example-Based Logical Labeling of Document Title Page Images" In: *9th Int. Conf. on Document Analysis and Recognition*, **2007**, p. 919—923

[3] H. Bunke, B. Messmer, "Similarity Measures for Structured Representations" In: *Topics in Case-Based Reasoning, First European Workshop, EWCBR-93,* Springer, **1994**, p. 837–849

[4] F. Cesarini, E. Francesconi, M. Gori, G. Soda, "A two-level knowledge approach for understanding documents of a multi–class domain." In: *5th Int. Conf. on Document Analysis and Recognition*, **1999**, p. 135–138

[5] F. Cesarini, E. Francesconi, M. Gori, G. Soda, "Analysis and understanding of multi-class invoices" In: *Int. Journal on Document Analysis and Recognition,* Springer, **2003**, p. 102–114.

[6] A. Dengel, B. Klein, "smartFIX: A Requirements-Driven System for Document Analysis and Understanding" In: D. Lopresti, J. Hu, R. Kashi (eds.), *Document Analysis V*, Springer **2002**, p. 433–444

[7] H. Hamza, Y. Belaid, A. Belaid, "Case-based reasoning for invoice analysis and recognition." In: *Case-Based Reasoning Research and Development,* Springer*,* **2007**, p. 404-418.

[8] H. Hamza, Y. Belaid, A. Belaid, "A Case-Based Reasoning Approach for Invoice Structure Extraction." In: 9*th Int. Conf. on Document Analysis and Recognition,* **2007**, p. 327–331.

[9] Insiders Technologies GmbH, Corporate Website: http://insiders-technologies.de

[10] B. Klein, A. Dengel, S. Agne, "Results of a Study on Invoice-Reading Systems in Germany" In: *Document Analysis Systems VI*, Springer, **2004**, p. 451—462

[11] B. Klein, A. Dengel, S. Agne, "On Benchmarking of Invoice Analysis Systems In: *Document Analysis Systems VII*, Springer, **2006**, p. 312—323

[12] D.P. Lopresti, G.T. Wilfong, "A Fast Technique for Comparing Graph Representations with Applications To Performance Evaluation" In: *Int. Journal on Document Analysis and Recognition,* Springer, **2004**, p. 219–229