A High-Performance Document Image Layout Analysis for Invoices

Conference Paper · April 2018

CITATIONS
2

READS
1,206

4 authors:

Mohammad Mohsin Reza
Technische Universität Kaiserslautern
3 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Md Ajraf Rakib
Ericsson
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Syed Saqib Bukhari
97 PUBLICATIONS **869** CITATIONS
SEE PROFILE

Andreas Dengel
Deutsches Forschungszentrum für Künstliche Intelligenz
699 PUBLICATIONS **6,144** CITATIONS
SEE PROFILE

A High-Performance Document Image Layout Analysis for Invoices

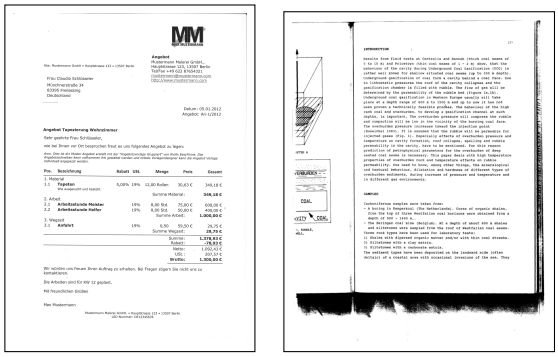
Mohammad Mohsin Reza*, Md. Ajraf Rakib*, Syed Saqib Bukhari, Andreas Dengel
DFKI and University of Kaiserslautern, Germany
{mohammad_mohsin.reza, md_ajraf.rakib, saqib.bukhari, andreas.dengel}@dfki.de

***Abstract*—Layout analysis for document is an important step in OCR pipeline and currently an intensive amount of research is going on to extract searchable full text from scanned images. Invoices are different in nature as compared to pages of books, magazine, loan documents and others, since, there are tables, header, footer, large white spaces, currency, item name, item amount, logo in the invoice. The standard layout analysis proves inefficient on invoices. In this paper we are proposing an advanced layout analysis for invoices that integrate the following steps in the standard layout analysis: removal of table cell lines and merging text lines. Additionally, we integrated the proposed layout analysis for invoices into the anyOCR system, which was mainly developed for both historical as well as contemporary documents from books, magazines etc. In the performance evaluation section, we will compare our advanced layout analysis pipeline with the standard anyOCR [1] pipeline and with a commercial OCR system like ABBYY. Our advanced layout analysis achieved better OCR accuracy as compared to the other mentioned systems.**

I. Introduction

There has been a resurgence of interest in optical character recognition (OCR) in recent years mainly for digitizing document to increase re-usability of information. Automatic data processing plays a vital role in processing lots of documents making our daily life not only easier but also get more benefit from the computerize system. A digital mailroom system is the automation of incoming mail (for example, scanned forms and invoices and digital emails) processes, where structured forms processing is relatively an easier task as compared semi-structured invoices. There are roughly two main tasks to process data from invoices: OCR and Information extraction. While performing end-to-end OCR pipeline for invoices, layout analysis in a most challenging task because of tables, header, footer, large white spaces, currency, item name and amount and logo, which are not commonly present in standard pages form books and magazines. These differences can be seen in Figure 1. Therefore, standard document layout analysis gives inefficient result on invoice as shown in evaluation.

In literature, there are some papers which proposed methods for document layout analysis. Bukhari et al. [2] proposed a layout analysis system for extracting Arabic text-lines from scanned documents written in different languages and styles. They presented the system based on a suitable combination of different well established techniques for analyzing Latin script documents that have proven to be robust against different types of document image degradations. In another paper, Bayer et al. [3] introduced a system that used an OCR tool with FRESCO model to extract particular information from invoices but they did not mention anything about the accuracies of that OCR tool as well as no detail about layout analysis techniques. Tuganbaev et al. [4] used a top-down document analysis structure with the FlaxiCapture technology to capture particular information from invoice. They performed a full-page OCR before applied their technology but also did not focus on the OCR accuracy. Though, the accuracy of the OCR tool is also important for extracting particular information so we were looking for some related paper that talks about some OCR method and its accuracy. But we did not find any particular paper that gives the better explanation for layout analysis of invoices.



(a) (b)

Fig. 1. (a) A sample scanned invoice, (b) A sample scanned standard (book page) document.

In this paper, we introduced an advanced high-performance layout analysis for invoices where we have integrated new methods for different tasks in the layout analysis pipeline of the anyOCR system [1], which is developed for processing pages from historical and contemporary books or magazines. For this purpose, we used a line removal method to remove line-graphics from the table and combined text lines so that information in the table are kept intact row by row. The rest of the paper is organized as follows. After discussing our proposed method in Section II for advance layout analysis of invoices followed by performance evaluation in Section III to compare our result with other systems. Finally in Section IV we conclude our work.

*These two authors contributed equally.

II. A High-Performance Layout Analysis for Invoices:

In order to understand our contribution in this paper, at first we will briefly describe the existing state of the anyOCR system and then we will present our contribution in its layout analysis pipeline.

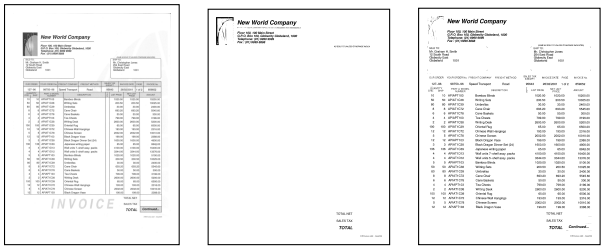
A. The anyOCR System - Overview [1]

The anyOCR component contains a set of document analysis methods that are usually required for a typical end-to-end OCR pipeline for extracting text from a document image. This method includes binarization, text and non-text segmentation, text line extraction, and producing OCR text in hOCR format, where text non-text segmentation and text line segmentation include as a layout analysis pipeline.

B. The Proposed anyOCR System for Invoices

We made some change in existing pipelines for layout analysis of invoices in the anyOCR system [1]. Usually, invoices contain table and most of the table draw with line-graphics that may recognize as a non-text part by existing pipeline which fails to extract text data from invoices. This is one of the first barriers for extracting data from invoices.

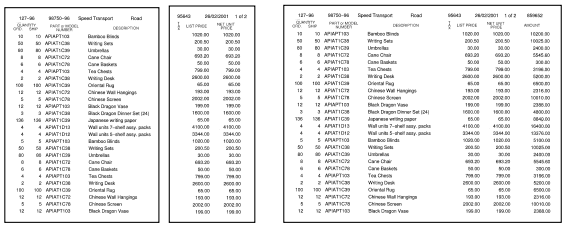
Firstly, we removed all line-graphics from invoices before applying binarization method. The results of an input image after processed by binarization and text non-text segmentation steps without and with our proposed line removal step are shown in Figure 2. One can see the results in this figure with line removal pipeline keep all the text in the image as compared to the existing pipeline.



(a) (b) (c)

Fig. 2. (a) A simple scanned invoice, (b) processing steps: binarization, text non-text segmentation, page frame segmentation, (c) processing steps: a newly introduced line-graphics removal, binarization, text non-text segmentation, page frame segmentation.

The existing text line segmentation method over-segments the columns of a table which may be difficult to synchronize with related data for each item in the table. We would like to intact them in a single line so that item information is extracted row by row which is shown in the Figure 3. Secondly, in order to achieve it, we changed the existing text line extraction method for invoice and merge the text lines segments with each other which are in the same row. Finally, text is recognized for each line in reading order and then saved in the hOCR format.



(a) (b)

Fig. 3. (a) Oversegmentation a table by standard anyOCR pipeline during page segmentation, (b) Our proposed method to keep intact row-wise information.

III. Performance Evaluation

There is no public dataset available for invoices. It is usually a laborious task to create datasets for performance evaluation. However, we have created a dataset of 29 images which we will partly share in public (because around 10 invoices are proprietary). For ground truth text creation, we considered all text entries at same height as a single text line. We compared the final OCR accuracy of the following systems: The standard anyOCR, ABBYY, and the proposed high-performance anyOCR system for invoices. The result are shown in Table 1, where our proposed system achieved the best performance as compared to the other two systems. ABBYY performed well on most of the images. However, for some cases, like Fig 2(a), it completely missed the whole table.

OCR System	Accuracy
The anyOCR system	53.95%
ABBYY	76.00%
The Proposed anyOCR system for Invoices	83.34%

TABLE I
The table Showing the OCR accuracy over three different pipelines including proposed one.

IV. Conclusion

The anyOCR is an open-source system which gives very good accuracy for standard document images such as pages from books, magazines and so on. Invoices are naturally different from standard document images because they contain tables, headers, footers. In this paper we integrated additional steps in the existing layout analysis pipeline of the anyOCR system to achieve a high-performance layout analysis for invoices. The proposed system achieved the best performance as compared to not only the existing anyOCR system but also the commercial system ABBYY.

References

[1] S. S. Bukhari, A. Kadi, J. M. Ayman, and A. Dengel, "anyocr: An open-source ocr system for historical archives," in *ICDAR*, 2017.

[2] S. S. Bukhari, F. Shafait, and T. M. Breuel, "High performance layout analysis of arabic and urdu document images," in *ICDAR*, 2011.

[3] T. Bayer and H. U. Mogg-Schneider, "A generic system for processing invoices," in *ICDAR*, 1997.

[4] D. Tuganbaev, A. Pakhchanian, and D. Deryagin, "Universal data capture technology from semi-structured forms," in *ICDAR*, 2005.