Analyzing the Potential of Layout Analysis Systems for the Task of Shopping Receipts Analysis on ReceiptDB Dataset

Shoaib Ahmed Siddiqui*†, Soumen Pramanik*†, Pervaiz Iqbal Khan*†, Andreas Dengel*†, Sheraz Ahmed†
*TU Kaiserslautern, Kaiserslautern, Germany
†German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
Email: firstname.lastname@dfki.de

***Abstract*—Shopping receipts are an important form of document which is ubiquitous as a proof of transaction all over the world. Despite recent advancements in the domain of document analysis, analysis of documents with complex layouts like receipts is still not a reality. In order to further investigate the efficacy of state-of-the-art layout analysis systems in the document processing community for the task of shopping receipt analysis, we curated a custom shopping receipts dataset comprising of 539 receipts collected from over 10 different supermarkets in Germany named as ReceiptDB. The dataset is densely labeled with the most important information which includes information regarding the row, product name, price, description, header, footer, logo, total price and total price text. Furthermore, in order to establish a baseline, we employed a state-of-the-art document analysis system powered by deformable FPN. It is evident from the obtained results that solving the problem of shopping receipt analysis will require significant efforts from the document analysis community in combination with the advances in deep learning literature.**

***Keywords*-Shopping Receipt Analysis, Layout Analysis, Document Understanding**

I. Introduction

With the rise in the use of technology, increasing emphasis is being laid on the digitization of old paper-based documents [1], [2]. The advantages of this digitization are threefold: (a) environmental friendly nature, (b) processing convenience and (c) reliable storage. Different solutions have been proposed in the past for the digitalization of these documents [2]. These documents (both ancient and recent) are comprised of literature, records, and other information like maps [1], [2].

Receipts are most commonly used proof for transactions all over the world. Managing these receipts manually is a big hassle for consumers willing to administer their monthly expenditures. Despite the advantages of this digitization, only a limited number of companies provide digital receipts. OCR systems have already achieved astonishing results for clearly visible and known text fonts [3], [4]. However, they still struggle in situations where the image quality is poor or the text is barely visible [5].

The most common way to analyze the layout of documents is based on the textual content [3]. Humans on the other hand, can directly analyze the layout without looking at the textual content present in the image. Therefore, this motivated us to answer an interesting question i.e. *Can we analyze the layout for shopping receipts by using state-of-the-art layout analysis systems using only the raw image as input without textual content?* In order to answer this, two main ingredients are required. The first ingredient is a large dataset for training and testing such a system. The second ingredient is the model itself.

We curated a custom receipt dataset comprising of data from over 18 different super-stores in Germany for training the receipt analysis system named as ReceiptDB. This dataset contains a total of 539 shopping receipts. The dataset demonstrates high-variability in layouts even for receipts from the same store. Some sample receipts are visualized in Fig. 1. To analyze the potential of layout analysis systems for shopping receipt analysis, we leveraged state-of-the-art deep learning based layout analysis system proposed by Siddiqui et al. (2018) [6]. The model utilizes the potential of deformable convolution operation to achieve state-of-the-art results on the task of table recognition in document images.

The rest of the paper is structured as follows. Section II highlights the previous work done in the domain of automatic analysis of shopping receipts. We then provide details regarding the self-collected dataset (ReceiptDB) in Section III. Section IV provides details regarding the presented approach for semantic analysis of receipt. Section V presents the obtained results along with a brief discussion. Finally, concluding remarks are presented in Section VI.

II. Related Work

Receipts analysis is a profoundly difficult problem as compared to other related tasks including table detection due to absence of any visual separators and poor receipt quality [6], [7]. Using the detected regions to extract the corresponding information adds yet another level of complexity into the system.

Generalized receipt analysis system which is able to generalize to arbitrary layouts is an audacious goal. Different efforts have been made in this direction for automatic analysis of user receipts due to its market potential. Altmeyer et al. (2016) [8] presented a thorough overview of the complexity of receipt analysis task along with the requirement for such a

(a) JYSK (b) MeinLaden (c) Douglas (d) Netto (e) Legend

Figure 1: Sample receipts from different companies

system. With supporting studies, they concluded that Smartphone application is the ideal way to realize such a system.

Zhu et al. (2007) [9] presented a solution for digitization of such receipts by using a combination of OCR, and Conditional Random Fields (CRF) for recognition of elements with high variance and regular expressions for recognition of elements with low variance. The system makes high errors for receipts not seen before. Receipts2Go [10] also used a regular expression based approach for extraction of elements with low variance along with usage of past OCR results to further improve their system. Thorough evaluation of their system is not presented raising a question towards its generalization capabilities. Shen et al. (2012) [11] relied on ontologies for extraction of entities from receipts using an object-relationship model, capturing information about sets of objects and their relationship sets. Their approach relied on perfect results from the OCR system making it ineffective in real-world conditions. Altmeyer et al. (2016) [8] presented a crowd sourcing based solution for correction of OCR errors utilizing game theory. The approach requires explicit human input making it infeasible to adapt for users.

Some of the most popular applications on Google PlayStore (Android) for receipt analysis are Smart Receipts[1] and Expensify[2]. Both apps provides the functionality for total amount detection. However, no product level information is extracted in both cases.

In all of these cases, the systems rely heavily on successful extraction of the textual content in order to successfully analyze the receipts. In contrast, we analyze the potential of state-of-the-art deep learning based document layout analysis system to directly analyze the layout of the shopping receipts similar to humans, based on just object locations, without having access to the actual textual content present in the shopping receipt.

[1]https://play.google.com/store/apps/details?id=wb.receipts

[2]https://play.google.com/store/apps/details?id=org.me.mobiexpensifyg

III. ReceiptDB

We present a new dataset which we call ReceiptDB. The labeling of images was carried out by LabelImg software[3]. Fig. 1 provides a few sample receipts from the dataset along with the provided annotations. The annotations are rich containing both the high level information like header, footer, rows etc. as well as fine-grained details like product name, price, description etc.

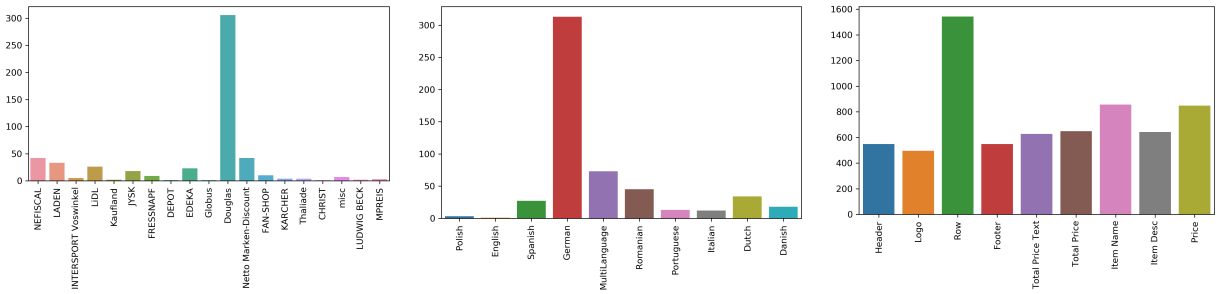
We selected a total of 9 classes which was abundantly present in the dataset. These classes include header, footer, logo, row (containing complete information regarding one product), produce name, product price, product description, total price and total price text. Row highlights all the information for one product, which can be considered analogous to high-level information. Product name, price and description on the other hand, provides the fine-grained information for every product inside the row.

The complete dataset distribution is visualized in Fig. 2. This includes distribution w.r.t. language, store and labels to provide a very detailed overview of the dataset. The receipts belong to more than 9 languages which are collected from more than 18 different super-stores. This provides a hint regarding the diversity of receipts present in ReceiptDB. Most of the receipts were collected from Douglas. Since the dataset was collected from different super-stores within Germany, the most common language was undoubtedly German. The distribution of the receipts according to the store in train, validation and test are presented in Table I. The segregation of languages into train, validation and test set is presented in Table II. It can be seen from the tables that the dataset is generated so as to maximize diversity.

IV. Method

We leverage the potential of state-of-the-art document layout analysis system proposed by Siddiqui et al. (2018) [6]. The system is comprised of a Feature-Pyramid Network

[3]https://github.com/tzutalin/labelImg



(a) Store Distribution (b) Language Distribution (c) Label Distribution

Figure 2: Dataset Distribution

Table I: Dataset distribution by store

Store Name	Train	Validation	Test	Total
CHRIST	1	0	0	1
DEPOT	0	0	1	1
Douglas	213	46	47	306
EDEKA	16	3	4	23
FAN-SHOP	6	2	2	10
FAN-FRESSNAPF	0	0	9	9
Globus	0	0	1	1
INTERSPORT Voswinkel	0	0	5	5
JYSK	12	3	3	18
KARCHER	4	0	0	4
Kaufland	0	0	2	2
MEIN-LADEN	23	5	5	33
LiDL	18	4	4	26
LUDWIG BECK	0	0	2	2
Miscellaneous	0	0	7	7
MPREIS	0	0	3	3
NEFISCAL	30	6	6	42
Netto Marken-Discount	29	6	7	42
Thaliade	4	0	0	4
Total	**356**	**75**	**108**	**539**

Table II: Dataset distribution by language

Language	Train	Validation	Test	Total
Danish	12	3	3	18
Dutch	23	3	8	34
English	1	0	0	1
German	210	39	64	313
Italian	8	2	2	12
Multi-Lingual	44	15	14	73
Polish	1	2	0	3
Portuguese	7	2	4	13
Romanian	30	5	10	45
Spanish	20	4	3	27
Total	**356**	**75**	**108**	**539**

(FPN) [12] equipped with the capabilities of the deformable convolution operation [13].

Convolution operation is at the core of state-of-the-art image classification and object detection models [14], [12]. The reason for its success over MLP methods is its parameter efficiency where a small filter is learned and reused at all the different locations in the image. Along with being parameter efficient, this also encourages the system to learn generalizable features, thus reducing over-fitting. Despite its success, the fixed receptive-field of the convolutional layers is problematic in cases where objects of interest vary significantly in terms of scale and pose.

In order to overcome this problem of fixed receptive field, Dai et al. (2017) proposed the deformable convolutional layers [13]. Deformable convolution operation generates explicit offset which allows the network to adapt its receptive field at every location in the image. These offsets are again computed by another set of convolutional layers which are also learned during backpropagation. This allows the network to mould its receptive field in order to cover it completely regardless of its scale and pose. The generated offsets are fractional, therefore, implemented via bi-linear interpolation [13].

Since offsets for the filter are generated at every location in the image, this is a very memory intensive operation. Therefore, only four layers at the top of the feature hierarchy are modified to reduce the memory footprint. Alongside these convolutional layers, the ROI-pooling layer is also equipped with the ability to deform its receptive field.

A common approach to cater for the detection of objects occurring at different scales is to do multi-scale detection. This requires forward-propagating the image through the network several times at different scales followed the aggregation of the results. The Feature-Pyramid Network (FPN) was proposed by Lin et al. (2017) [12], capable of directly detecting multi-scale objects with only a single forward-pass through the network by adding top-down pathways performing inference at different resolutions and levels of abstraction.

The deformable FPN [13] combines the capabilities of both deformable convolution operation and FPN to obtain the best of both worlds. The model achieved the best performance on the task of table detection in document images [6]. The deformable FPN uses a deformable ResNet-101 base model (with 4 deformable layers) and deformable position-sensitive ROI-pooling layer.

In order to reap the advantages of transfer learning, we used the ImageNet pretrained ResNet-101. Training on such a large dataset transforms the initial layers of the network to generalized feature extractor, while the layers at the end (mostly dense layers) are task specific. We reuse the convolutional layers of the network. This significantly boosts

the rate of convergence, along with reduction in over-fitting in many cases where the dataset in hand is small which was the case for ReceiptDB.

V. Results

The results on ReceiptDB are presented in Table III. We report results from the deformable FPN. Alongside, we also report results from deformable F-RCNN for comparison. The results from deformable Faster R-CNN and deformable R-FCN were very similar, therefore, we only report the results from deformable F-RCNN. This is consistent with the findings of Siddiqui et al. (2018) [6].

We report the commonly used object detection metrics which includes precision, recall and F-Measure. In the document analysis community, two alternate reporting schemes are used for reporting the results. The first scheme first computes the True Positives (TP), False Positives (FP) and False Negatives (FN) over the entire test set and accumulates them. With these accumulated TP, FP and FN numbers, the final precision, recall and F-Measure are computed. The second scheme follows per-document averaging, where the precision, recall and F-Measure is computed for every document based on the TP, FP and FN which is then averaged over the complete test set. The first scheme is useful in providing an overall picture of the system by taking every TP, FP, and FN into account. The second scheme avoids being biased towards any particular document containing a particularly large number of objects. Therefore, both metrics provides a different view of the system. For the sake of brevity, we report the number from both these schemes where the first scheme is highlighted as *Accumulated Scores* in Table III while the second scheme is named as *Per-Document Average Scores*.

It is evident from the results that the system was able to correctly detected the header, footer and logo in most of the receipts, pertaining to a high F-Measure ($> 90\%$). Information like row, total price and total price text which was fairly consistent among the different receipts was also correctly segmented in most of the cases, also resulting in a reasonable F-Measure ($> 80\%$). Fine-grained information like item name, item price and item description was difficult for the network to segment out, resulting in low F-Measure ($\sim 70\%$). It is interesting to note that the price is usually the easiest information to segment out within a row, usually located on the right-side of the receipt with a fairly consistent notation across the different stores. However, the markings of the name and price were highly subjective in some cases, requiring particular knowledge regarding the product names. This resulted in significantly lower F-Measures for the two fine-grained details. The per-document average doesn't provide a real picture in this case since the system performed poorly on all the cases where there were a large number of products, however, the number of receipts containing a large number of products were themselves quite rare. The F-Measure computed by accumulating scores for item name and item description is very low ($\sim 50\%$).

Some sample correct examples are visualized in Fig. 3. It can be seen from the results that almost all of the perfectly recognized images are from Douglas since Douglas was the most common receipt in the dataset alongside the simple layout. One fan-shop receipt was also perfectly recognized as there was no product-level information available in the receipt, making the recognition process easy.

Some incorrect examples are visualized in Fig. 4. In cases where similar receipts were present in the dataset, the system did a reasonable job in detecting the relevant items (Fig. 4a and Fig. 4b). On the other hand, the system demonstrated poor generalization capabilities on receipts with a novel layout or with significant number of products (Fig. 4c, Fig. 4d and Fig. 4e). This could be in part because of the poor product level resolution upon resizing while maintaining aspect ratio in order to avoid exhausting the GPU memory.

VI. Conclusion

This paper a presented a new problem of recognizing the layout of the receipt without explicitly looking at the textual content. In order to investigate the problem, a new dataset is curated comprising of 539 receipts collected from more than 10 different super-stores in Germany encouraging diversity. A state-of-the-art deep learning based layout analysis system was utilized for the task (deformable FPN). The system achieved high performance for segmentation of high-level information like header, footer, and logo. However, struggled significantly in segmenting-out the fine-grained information. The obtained results advocate that the problem of receipt layout analysis is still far from reality despite all the recent advancements in the domain of deep learning and document analysis. Significant efforts needs to be invested from the document analysis community in order to solve the layout analysis problem with such high diversity and complexity.

Acknowledgment

Hidden for double blind review.

References

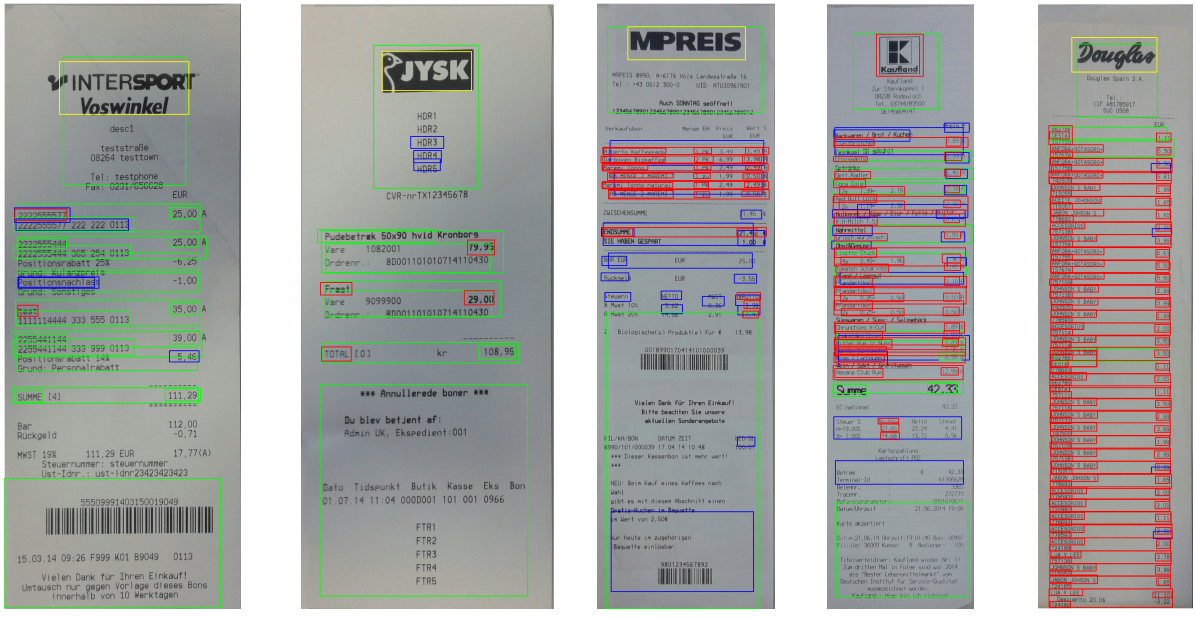
[1] L. Eikvil, K. Aas, and H. Koren, "Tools for interactive map conversion and vectorization," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 2. IEEE, 1995, pp. 927–930.

[2] F. Kleber, M. Diem, F. Hollaus, and S. Fiel, "Mass digitization of archival documents using mobile phones," in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, ser. HIP2017. New York, NY, USA: ACM, 2017, pp. 65–70. [Online]. Available: http://doi.acm.org/10.1145/3151509.3151526

[3] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

(a) (b) (c) (d)

Figure 3: Correctly Recognized Tabular Structures. Yellow color depicts TP for log while Green color depicts TP for the rest of the classes.



(a) (b) (c) (d) (e)

Figure 4: Incorrectly Recognized Tabular Structures. Green color depicts true positives for rows, yellow color depicts true positives for columns, red color indicates false negatives for both rows and columns, blue color depicts false positives for rows, and magenta color indicates false positives for columns.

					Accumulated Scores			Per-Document Average Scores		
Models	**Class**	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Deformable FPN	Total Price	120	41	16	0.745342	0.882353	0.808081	0.886883	0.916667	0.862397
	Footer	101	13	7	0.885965	0.935185	0.909910	0.912037	0.939815	0.919753
	Item Name	103	61	106	0.628049	0.492823	0.552279	0.762654	0.780864	0.718331
	Price	115	21	91	0.845588	0.558252	0.672515	0.895525	0.855703	0.828520
	Header	104	3	3	0.971963	0.971963	0.971963	0.972222	0.972222	0.962963
	Item Description	63	30	91	0.677419	0.409091	0.510121	0.848765	0.726852	0.701587
	Logo	94	15	4	0.862385	0.959184	0.908213	0.929012	0.967593	0.927469
	Total Price Text	125	25	16	0.821429	0.877863	0.848708	0.893827	0.895833	0.855467
	Row	279	69	76	0.801724	0.785915	0.793741	0.872451	0.867797	0.847376
Deformable FRCNN	Total Price	93	73	43	0.560241	0.683824	0.615894	0.711883	0.772377	0.678968
	Footer	104	7	4	0.936937	0.962963	0.949772	0.958333	0.967593	0.953704
	Item Name	66	74	143	0.471429	0.315789	0.378223	0.676235	0.585494	0.516578
	Price	90	48	116	0.652174	0.436893	0.523256	0.764198	0.693313	0.626455
	Header	104	4	3	0.962963	0.971963	0.967442	0.962963	0.972222	0.962963
	Item Description	45	23	109	0.661765	0.292208	0.405405	0.805556	0.615741	0.592593
	Logo	93	2	5	0.978947	0.948980	0.963731	0.986111	0.958333	0.956790
	Total Price Text	92	71	39	0.564417	0.702290	0.625850	0.785979	0.736883	0.675112
	Row	230	94	125	0.709877	0.647887	0.677467	0.759768	0.792635	0.755004

Table III: Results on ReceiptDB

[4] T. M. Breuel, "The ocropus open source ocr system," in *Document Recognition and Retrieval XV*, vol. 6815. International Society for Optics and Photonics, 2008, p. 68150F.

[5] G. Naganjaneyulu, A. Narasimhadhan, and K. Venkatesh, "Performance evaluation of ocr on poor resolution text document images using different pre processing steps," in *TENCON 2014-2014 IEEE Region 10 Conference*. IEEE, 2014, pp. 1–4.

[6] S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, and S. Ahmed, "Decnt: Deep deformable cnn for table detection," *IEEE Access*, vol. 6, pp. 74 151–74 161, 2018.

[7] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "Deepdesrt: Deep learning for detection and structure recognition of tables in document images," in *ICDAR*, vol. 1. IEEE, 2017, pp. 1162–1167.

[8] M. Altmeyer, P. Lessel, and A. Krüger, "Expense control: a gamified, semi-automated, crowd-based approach for receipt capturing," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 2016, pp. 31–42.

[9] G. Zhu, T. J. Bethea, and V. Krishna, "Extracting relevant named entities for automated expense reimbursement," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 1004–1012.

[10] B. Janssen, E. Saund, E. Bier, P. Wall, and M. A. Sprague, "Receipts2go: the big world of small documents," in *Proceedings of the 2012 ACM symposium on Document engineering*. ACM, 2012, pp. 121–124.

[11] Z. Shen and Y. Tijerino, "Ontology-based automatic receipt accounting system," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03*. IEEE Computer Society, 2012, pp. 236–239.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE CVPR*, 2017, pp. 2117–2125.

[13] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," *CoRR, abs/1703.06211*, vol. 1, no. 2, p. 3, 2017.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.