TSR-DSAW: Table Structure Recognition via Deep Spatial Association of Words

Arushi Jain, Shubham Paliwal, Monika Sharma, Lovekesh Vig

TCS Research, Delhi, India
Email:{j.arushi, shubham.p3, monika.sharma1, lovekesh.vig}@tcs.com

Abstract. Existing methods for Table Structure Recognition (TSR) from camera-captured or scanned documents perform poorly on complex-tables consisting of nested rows / columns, multi-line texts and missing cell data. This is because current data-driven methods work by simply training deep models on large volumes of data and fail to generalize when an unseen table structure is encountered. In this paper, we propose to train a deep network to capture the spatial associations between different word pairs present in the table image for unravelling the table structure. We present an end-to-end pipeline, named *TSR-DSAW: TSR via Deep Spatial Association of Words*, which outputs a digital representation of a table image in a structured format such as HTML. Given a table image as input, the proposed method begins with the detection of all the words present in the image using a text-detection network like CRAFT which is followed by the generation of word-pairs using dynamic programming. These word-pairs are highlighted in individual images and subsequently, fed into a DenseNet-121 classifier trained to capture spatial associations such as same-row, same-column, same-cell or none. Finally, we perform post-processing on the classifier output to generate the table structure in HTML format. We evaluate our TSR-DSAW pipeline on two public table-image datasets - PubTabNet and ICDAR 2013, and demonstrate improvement over previous methods such as TableNet and DeepDeSRT.

1 Introduction

An important bottleneck in the digitization of camera captured or scanned documents such as invoices, resumes, and insurance forms is the presence of tabular data in an unstructured format. This tabular data is required to be digitized and stored in a machine understandable format like HTML/XML to fully capture the layout and logical structure of the table. Manual digitization of a large number of such document images containing tabular data is time-consuming, labour-intensive and error-prone. Hence, there is an urgent need to automate the digitization process using Table Structure Recognition (TSR) techniques. However, diversity in table layouts, complexity of nesting (spanning multiple rows and columns), multi-line texts, missing cells and overlapping columns, have made TSR a challenging task.

Recent methods such as TableNet [1], DeepDeSRT [2], Graph Neural Networks based table recognition [3] often fail to work on very complex tables due to their low generalization ability. Moreover, these methods rely on very complex and deep neural network architectures which require huge amount of data

257

and computational resources for their training. In addition, they ignore the implicit information available in the table images in the form of spatial association between different words which can prove to be very beneficial for capturing the table structure efficiently. This motivates us to propose a novel and simple end-to-end pipeline for TSR which takes a table image as input and returns its structural representation in HTML format by utilizing the spatial association between different words. Please note that we focus on extracting table structure information given an image containing table only.

The proposed method of TSR using deep spatial association of words (TSR-DSAW) is a combination of deep learning and dynamic programming techniques. It works by first detecting all the words present in the given table image using a text-detection module such as CRAFT (Character Region Awareness for Text Detection) network [4], followed by forming relevant word-pairs efficiently using dynamic programming technique and subsequently, determining the associations between these word-pairs using a DenseNet-121 based Word2Word Association classifier e.g., same-row, same-column, same-cell or none. At the end, we generate HTML structure of the input table by applying post-processing on the outputs of word-association classifier. The proposed TSR-DSAW method is simple, robust and computationally efficient.

2   Related Work

Heuristics Based Methods: Earlier conventional approaches [5, 6, 7, 8] require extensive amount of human-effort for writing complex algorithms based on text-block arrangement and grid-patterns for recognizing the table layout. These methods are unable to handle the diverse layouts of the tables and the variety of grid-patterns, missing cells, row and column spans in real world datasets.
Deep Learning Methods: These are data-driven approaches which do not make any prior assumption about table structures and leverage deep-networks that are trained on large amount of data. One such method [2] trains a deep learning network on Marmot dataset [9] for locating structural components like rows, columns and cells. Later, Paliwal et al. [1] proposed an end-to-end deep learning based method called TableNet for detecting table and columns by a single network where they used a single encoder and two decoders, one for each detection task. Authors in [10] demonstrate the image-to-markup model which was trained on a big TableBank dataset to predict XML from the table image directly. Though these methods work well for simple tables, they do not produce robust performance on nested tables, and tables with overlapping columns.
Graph Based Methods: Recently, many researchers have formulated the table recognition problem as a graph, mapping words to nodes and edges to relationships between word pairs. Qasim et al. [3] propose a solution with Graph Neural Networks to find the association between words on a synthetic dataset of tables. Although, they have shown remarkably good results for synthetic tables, the method does not translate well to real tables which are far more diverse in their structure patterns. In another paper [11], authors combine the rule based

approach with graph models where they propose an algorithm called RAC (Remove and Conquer) for building layout regions and then, use a graph model to describe their arrangement. Hu et al. [12] introduce a Directed Acyclic Graph (DAG) based model for hierarchical clustering to define columns and headers that are identified using spatial and lexical criteria. Very recently, [13] proposed attention based solution called GraphTSR which uses a transformer to predict the adjacency relationship of individual cells. This method made significant improvement in understanding complex tables assuming the availability of accurate bounding boxes and digital content of individual cells.

3 Proposed Method: TSR-DSAW

TSR-DSAW takes a table image as input and comprises of 4 modules, as shown in Figure 1: (a) extraction of individual words; (b) Word-pair generation; (c) determination of Word2Word association; and (d) post-processing to generate structured output in HTML format.
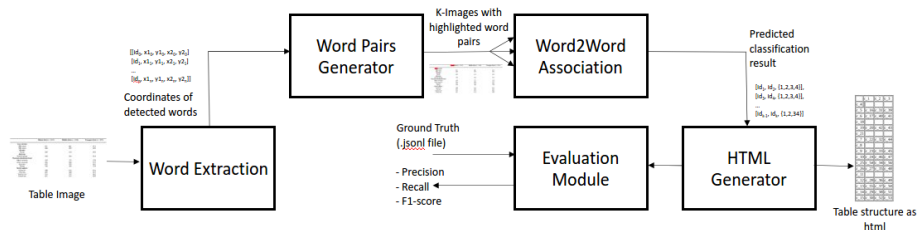

Fig. 1: The proposed TSR-DSAW pipeline for extracting table structure

Word Extraction: Initially, we pre-process the input image to handle alignment issues using Radon transformation as mentioned in TableNet [1]. The pre-processed image is divided into overlapping patches of size 512×512 having 50% overlap with adjacent patches. These patches are then passed through CRAFT [4] which localizes individual text-regions and gives us their bounding boxes. Next, the bounding box coordinates of words present in the overlapped patches are merged which may lead to some redundancy in scenarios where the overlaps span multiple rows or columns. To circumvent this, we scan the entire image in small frames and remove the biggest box where multiple boxes cover a single frame. Finally, the extracted text-regions are read using Tesseract[1].

Word Pair Generation: One trivial approach for generating word pairs is that given a word in a collection of w words, make pairs with the remaining $(w - 1)$ words leading to $O(w^2)$ complexity. However, this approach is inefficient. Therefore, we propose a dynamic programming based word-pair generation technique

---

[1]Tesseract v4.1.1: https://github.com/tesseract-ocr/tesseract

with a linear time-complexity of $O((m + n)w)$ and claim that we need a maximum of $m + n$ combinations for a word, where m and n are the number of rows and columns in a table. We support the correctness of our claim with the explanation that only the immediate left neighbors and top neighbors need evaluation for nested-rows and nested-columns, respectively. To handle the diversity in the size of words, we capture all the immediate words which are its top and left vicinity. We can approximate the value of m and n based on maximum number of text-regions to be considered in both the directions. In our experiments, we found 3 to be an optimal value for both parameters.

Word2Word Association: This module uses a DenseNet121 based classifier to exploit the spatial relationship between different word-pairs such as 'Same Row', 'Same Column', 'Same Cell' and 'None' for identifying table-structure. In order to train the Word2Word Association classifier, we used 3600 tables randomly selected from train-set of PubTabNet dataset to obtain input images of size 224×224 having highlighted word-pairs. The solid-red color boxes are used to represent highlighted word-pairs which also eliminates the OCR and language dependency from our model. This allows the network to entirely focus on the spatial bounds and hence, learn the spatial relationships between them. We generated $48k$ such input images covering nested/harder cases along with simpler cases. Nested cases are the word-pairs which are related with "Same Row/Column" association but structure is nested in terms of multiple-spans. On the other hand, harder cases are those pairs that are difficult to infer such as, pairs which are quite far away but are in same cell or are near to each other but are not related in anyway. An important point to note here is that the proportion of simple and hard cases is kept equal in the training-set.

HTML Generation: This step is highly dependent on the Word2Word Association classifier results as a single wrong classification can return an entirely different structure. Assuming that spatial associations are correct, a directed graph is generated where nodes are cells and edge $u \rightarrow v$ defines u and v are in 'Same Row' and u tends to be at left of v in spatial bounds. Similarly, another graph for column information is created where edge represents u to be above v. Graphs are traversed recursively in bottom-up fashion to collect spanning information. Row/col span of a node is defined by the sum of row/col span of its child nodes which are reachable by a single path. If a node has no child, the span is set to 1. Using this spanning and row/column child information from the graphs, a $m \times n$ matrix is created, where m and n defines number of rows and columns respectively and the value in the matrix defines the table cell (denoted by an identifier). The resultant matrix is then converted into HTML format.

4   Experiments

Dataset Details: As stated earlier, we use a subset of table images from Pub-TabNet train-set for training of Word2Word association classifier. For evaluation of TSR-DSAW, we select 4800 table images randomly from val-set of PubTab-

Table 1: Performance Comparison of TSR-DSAW against TableNet [1] and DeepDeSRT [2]. Please note that Acc_{w2w} refers to the classification accuracy of Word2Word association classifier.

Method	TestSet	Acc_{w2w}	Precision	Recall	F1-Score
TSR-DSAW	PubTabNet	0.9287	**0.9625**	**0.9087**	**0.9348**
TableNet	PubTabNet	-	0.9572	0.8788	0.9163
TSR-DSAW	ICDAR13	0.9215	**0.9649**	**0.9195**	**0.9416**
TableNet	ICDAR13	-	0.9255	0.8994	0.9122
DeepDeSRT	ICDAR13	-	0.9593	0.8736	0.9144

Net dataset because PubTabNet has not released ground-truth for its test-set. In addition, we perform evaluation on ICDAR13 [14] which includes 152 table images extracted from European Union and US PDF reports.

Evaluation Metrics: The evaluation of table structure is done based on adjacency relations of two cell structures [15, 16]. In order to avoid the errors due to OCR, we replace the cell content and corresponding ground truth with a unique ID. We compute adjacency relations with nearest neighbors in horizontal and vertical directions. No links are created when either of the cell is blank. Next, we compute metrics [14] such as precision, recall and f1-score in order to compare the adjacency relations between predicted and ground-truth table structure.

Experimental Results: Now, we present comparison results of TSR-DSAW against two previous methods of TSR - TableNet [1] and DeepDeSRT [2]. It is evident from Table 1 that TSR-DSAW shows significant improvement over existing methods in all the metrics on both the test datasets. For instance, in case of PubTabNet test-set, our method TSR-DSAW obtains a higher F1-score of 0.9348 as compared to TableNet (0.9163). Similarly, TSR-DSAW beats both TableNet and DeepDeSRT on ICDAR 2013 test-set by a considerable margin and achieves 0.9649, 0.9195 and 0.9416 as precision, recall and F1-score, respectively. An important thing to note here is that our Word2Word association classifier is trained only on images from PubTabNet train-set as mentioned in Section 3 and is evaluated on the ICDAR 2013 dataset without any further fine-tuning. Further, the classification accuracies of Word2Word association are 0.9287 and 0.9215 on PubTabNet and ICDAR 2013 test-sets, respectively. Due to page limit constraints, qualitative results can be found here [2].

5 Conclusion and Future Work

We presented a deep-learning based simple method called TSR-DSAW which leverages the spatial association between words present in the table image to uncover table structure and produces an output in digital format such as HTML. The proposed method is robust in recognizing complex table structures having multi-span rows/columns and missing cells. The deep-learning based Word2Word

---

[2]Qualitative Results - TSR-DSAW: https://github.com/arushijain45/TSR-DSAW

association classifier is trained on a subset of PubTabNet train-set where it uses solid color boxes to highlight word-pairs which makes it language-independent. The proposed method is evaluated on PubTabNet validation set and ICDAR 2013 dataset without any fine-tuning. We outperform existing methods of TSR such as TableNet and DeepDeSRT on these datasets. In future, we would like to further improve the performance of Word2Word association classifier by incorporating more diverse table data and experimenting with other deep networks such as vision transformers.

References

[1] Shubham Paliwal, D. Vishwanath, R. Rahul, Monika Sharma, and L. Vig. TableNet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. *ICDAR*, pages 128–133, 2019.

[2] 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*, 2017.

[3] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. Rethinking table parsing using graph neural networks. *CoRR*, abs/1905.13391, 2019.

[4] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9357–9366. IEEE Computer Society, 2019.

[5] Richard Zanibbi, Dorothea Blostein, and James Cordy. A survey of table recognition. *IJDAR*, 7:1–16, 03 2004.

[6] Basilios Gatos, Dimitrios Danatsas, Ioannis Pratikakis, and Stavros J. Perantonis. Automatic table detection in document images. In Sameer Singh, Maneesha Singh, Chid Apte, and Petra Perner, editors, *Pattern Recognition and Data Mining*, pages 609–618, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[7] Scott Tupaj, Zhongwen Shi, C. Hwa Chang, and Dr. C. Hwa Chang. Extracting tabular information from text files. In *EECS Department, Tufts University*, 1996.

[8] Gaurav Harit and Anukriti Bansal. Table detection in document images using header and trailer patterns. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '12, New York, NY, USA, 2012. Association for Computing Machinery.

[9] Jing Fang, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 445–449, 2012.

[10] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, M. Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In *LREC*, 2020.

[11] E. Koci, M. Thiele, W. Lehner, and O. Romero. Table recognition in spreadsheets via a graph representation. In *2018 13th IAPR Workshop DAS*, 2018.

[12] Jianying Hu, Ramanujan S. Kashi, Daniel P. Lopresti, and Gordon Wilfong. Table structure recognition and its evaluation. In *Document Recognition and Retrieval VIII*, volume 4307, pages 44 – 55. International Society for Optics and Photonics, SPIE, 2000.

[13] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xianling Mao. Complicated table structure recognition. *CoRR*, abs/1908.04729, 2019.

[14] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In *ICDAR*, 2013.

[15] Matthew Hurst. A constraint-based approach to table structure derivation. In *ICDAR*, USA, 2003. IEEE Computer Society.

[16] Max Göbel, Tamir Hassan, and et al. A methodology for evaluating algorithms for table understanding in pdf documents. In *ACM Symposium on Documents*, USA, 2012.