International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Survey on Table Detection from Document Image

Shashank Jain[1*], Amritesh Singh[1], Rahul Ranjan Singh[1]

[1]Master Student, Dept. of Master of Computer Application, Jain Deemed to be University, Bengaluru, India

---

Abstract - *There are many types of invoice having table exist in the current system such as table in native text invoices, table in image invoices (II), table in handwritten invoices (HI) and so on. Nowadays, these different types of invoices are processing manually. Now our aim to survey such system which can handle invoices having the table automatically by using OCR (Optical Character Reader) and Deep Learning Technologies. Moreover, we will also discussed multiple technologies and suggest the best model as per our survey***.**

***Key Words*: OCR, Table Detection, Table Structure, Table Segmentation, Hand-written table detection.**

1. INTRODUCTION

Invoices has its different kind of format having a different style of table. The most important technique before processing for OCR needs to improve the quality of image invoice. This OCR information is more important to process invoices automatically to extract the valuable information by applying the indexing to get faster-searching result. The author also shared about need a different OCR approach than II [1]. Author produce an approach to improve the computer vision problem for getting a better result and also he produced two approach. 1) Faster Recurrent Conventional Neural Network (F-RCNN) [2] [3] algorithm to detect the table. 2) Region Proposal Network (RPN) to use transformed images as an input and return output set of rectangular proposal [4]. The Author produced an approach to detect table headers for the multi-dimensional table with the help of the Machine Learning Classification algorithm by using the two-sample dataset from CiteSeer documents repository. [5]**.**

2. LITERATURE SURVEY

Naganjaneyulu, et al. [6]**,** proposed a model on the multi clue heuristic based algorithm for table detection. He also proposed an approach used two parallel process Hough line and Hough corner point both is based on multi clue heuristic algorithm. The author also used different types of document having tables to observe accuracy of 89%. The author also present his theory on: 1) Hough line (to detect the shape of image like circle, parabola etc.). 2) Hough corner detection (used to detect the corner of image which can be defined having intersection of two edges).

Advantages

1. The author used a filter in image to enhance the image resolution based on weight sobolev space (WSS) [7] parameter $\alpha = 0.3$ to reduce the breakage.
2. He used Hough line and Hough corner parallel process based on heuristic method to detect the corner of tables.
3. The proposed method is work well with tables having general rectangular box layout.
4. His method gave the evaluation result with accuracy of 89% of his own created dataset.

Disadvantages

1. His method fails with tables having with space delimitation and irregular table format.
2. The cost of his proposed method is very expensive.

Dhiran, et al. [8]**,** proposed to this article to detect and extract the table information from image document where he used OCR for preprocessing which involve binarization, noisy border removal and enhancements. The author used 1200 image document which contains all types of format shown in (Tables 1). In this article, the author categories the table into three types –

1. Binarization

The author used this process for converting color image to binary image to make image quality better.

2. Noisy Border Removal

The author used this process for converting color image to binary image to make image quality better.

3. Enhancement

For enhancement, he applied a dilation process in this process it is identify small gaps to remove which make it noisy and structure the elements in rectangular box by joining the lines. He also used to white space to segment text line which occurs between text-line. The segmentation calculation is done by calculating the

horizontal (for row) segment and vertical (for column) segment.

Type	Description
Type 1	Has line row and column separator.
Type 2	Has line row and space for column separator.
Type 3	Has space for both row and column separator.

Table 1: Types of Table Format

Advantages

1. It used binarization, and noisy border removal process to make image quality better.
2. He used dataset to test having all three types of table format.
3. It gives 88.7 % accurate result on all three types of tables patterns.
4. His proposed method works well to detect on table Type 2 & Type 3 while [6] proposed method fails.

Disadvantages

1. In some test case, connect components behaves like table detected.
2. His proposed method will not work to detect table from multiple column page document.

Shafait, et al. [9]**,** proposed his approach to detect tables from the heterogeneous documents having a various layout of table. He also proposed an algorithm to detect the table with high accuracy on documents having a various types of layout table. He also used an open source Tesseract OCR and to evolution of algorithm, he used UNLV open source dataset. He also proposed his approach to analysis on documents via tab-stop (is a location where text aligns like top, left, right, center etc.) detection in various steps –

1. First step is to pre-process the document image to identify the vertical and horizontal line (separator) and locate image regions in documents. After that do the analysis on connected components to identify the candidate text components based on their size and stroke width.
2. The candidate dates are merged together into vertical line to search tab-stop position that are aligned vertically.
3. The column layout of page is inferred and connected components are merged into column position partitions is a sequence of connected components that don't cross any table line.
4. In the last step, the author creates a flow of the column partitions. He just grouped all font-size and line spacing into different blocks of partition columns.
5. He proposed his approach of table spotting on table detection algorithm which is built upon two components of layout analysis module: 1) Column Partitions - Column partitions give us connected components grouped by their type into partitions that do not cross tab-stop lines. 2) Column Layout – The layout analysis is done in presence of table regions. Based on the above analysis, the table detection algorithm is designed –

   i. **Identify Table Partitions**
   The author finds the text column partition which could belongs to table regions referred as table partitions. He also proposed three types of partition based on his observations: a.) partitions should have at least one more gap amid their connected components. b.) Partition having only single word. c.) partitions that overlaps with y-axis.

   ii. **Detecting Page Columns Split**
   This method occurs when the cell of the tables is well aligned. He divide the page into columns and calculate table's ratio partition in each column to detect the table.

   iii. **Locating Table Columns –**
   The main aim for this step is to combine all table partitions into a table.

   iv. **Marking Table Regions**
   It marked in rectangular box after identifying table regions.

   v. **Removing False Alarm**
   False alarms consisting of single column are removed by analysing their projection on the x-axis. Projection of a valid table on the x-axis should have at least one zero-valley larger than the median x-height of page.

Advantages

1. He did layout analysis to identify the column layout, column partition and segmentation block.
2. His proposed model also work on tab-stop detection for layout analysis.
3. His proposed model can also detect the table region.

Disadvantages

1. His proposed has no feature of table structure extraction.
2. His proposed model does not work on hand written table.

Huynh-Van, et al. [10]**,** proposed a hybrid method which has steps to detect the table zones: 1) region classification

2) detect table having intersection of vertical and horizontal lines and 3) identification of table made up by only parallel lines. He is used dataset UW-III for experiment to obtain a result. The author proposed his method in several steps:

1. Region Classification

In this process, first consider the large connected components is larger if height and width larger than 2% of image height and 6% of image width respectively. After that he identify the region which belongs to the convex hull of large connected components consider has table region.

2. Features for detecting a table containing intersection lines

An analysis of connected components on the intersection line is used to extract features of tables.

3. Features for detecting a table containing horizontal and vertical lines

In this proposed method, we also consider the table regions that are created by one or more horizontal/vertical line in region. The vertical/horizontal alignment text lines in this region are grouped together. Based on this group, local features and contextual features are extracted to validate it belongs to a table or not.

Advantages

1. The proposed method used region classification algorithm for both types of tables.
2. He used also local, contextual and combine feature for detecting table having a horizontal line and vertical lines.
3. In his proposed method algorithm can also detect the vertical tables.
4. He did evaluation on dataset UW-III with Abbey OCR, Tesseract OCR and Proposed Method based on text based and ratio based on parameters correct, partial, under, over, missed, and false.
5. He got the accuracy of precision (82%) and recall (80%).
6. His proposed method takes approx. 1 second per document images.

Disadvantages

1. His proposed method do not work on table having white space.
2. If having a large number of document then the complexity and computation time can vary.

Gilani, et al. [4], proposed a method which has two module Image Transformation and table detection. The author suggest in his article to use Abbey Cloud OCR SDK and tesseract OCR to show comparison result.

1. Image Transformation

In this module, image processing plays the most epochal role in natural image so we can easily fine-tune by applying on the existing F-RCNN model. He also proposed distance transformation to calculate the exact distance between the text region and white space present in image documents to get a better estimation about the existence of the table region by deriving representation of image document. In this proposed model, the author used different types of distance transformation so that can store different feature in three channels. The author performed an image transformation by following some procedure as mentioned below (Table 2).

Table 2: Image Transformation Procedure (I)

Parameters	Functions
b	Euclidean Distance Transformation (I)
g	Linear Distance Transformation (I)
r	Max Distance Transformation (I)
p	Channel Merge (b, g, r)

2. Table Row Detection

After table matching and baseline detection, he proposed to detect the row from table. First, the hand-written line do not consider for delimiting the rows. Once the text-line and column identified then each text-line will be tagged in B (Beginning), I (Inside), E (End), S (Singleton), O (outside of table) which correspond of position of text-line in cell. BIESO is an extension of Natural Language Processing is used to recognize the sequence of words in a sentence. Two graph-based approaches for categorizing text lines into BIESO categories are compared: Conditional Random Fields (CRF) and Graph Convolutional Networks (GCN).

Advantages

1. His proposed model is based on handwritten table detection.
2. To match the relational structure, he used maximum clique in an auxiliary graph structure (association graph) to solve.
3. He used a kerbosch algorithm to find the maximal clique in undirected graph.
4. He used BIESO based on Natural Languages Processing is used to recognize the word from sentence.

Disadvantages

1. His proposed model does not work on table not having row and column line.

Kleber, et al. [11], proposed his approach a template based table structure matching using where he applied two method to recognize table. The evaluation of methodology is done on the historical register book (death record) of the archive of the diocese of Passau. The author proposed to detect structure of handwritten/printed tables on dataset of ABP_S_1847-1878 provided by Passau Diocesan Archive having a total of 25,579 scanned pages. The author proposed his table recognition method in two parts –

1. Table Structure Matching Using Association Graph

The proposed methodology match the given template's table structure on the visible line information. The table structure of template represent as extends of PAGE XML which defines the logical and physical representation of documents. The documents comprises of table headers and columns. The author proposed a solution to match the relation structure by transforming it into the equivalent problem of finding the maximum clique in an auxiliary graph structure known as association graph. He also proposed to find the maximal clique in unidirectional graph by Kerbosch algorithm where clique is a subset of vertices. In this methodology, at the first the rough alignment of table is determined by the correlation of template image and documents. The author also define in his article how to detect and match based on line model for each table cell and resulting alignment error using the association graph algorithm.

2. Table Row Detection

After table matching and baseline detection, he proposed to detect the row from table. First, the hand-written line do not consider for delimiting the rows. Once the text-line and column identified then each text-line will be tagged in B (Beginning), I (Inside), E (End), S (Singleton), O (outside of table) which correspond of position of text-line in cell. BIESO is an extension of Natural Language Processing is used to recognize the sequence of words in a sentence. Two graph-based approaches for categorizing text lines into BIESO categories are compared: Conditional Random Fields (CRF) and Graph Convolutional Networks (GCN).

Advantages

1. His proposed model is based on handwritten table detection.
2. To match the relational structure, he used maximum clique in an auxiliary graph structure (association graph) to solve
3. He used a kerbosch algorithm to find the maximal clique in undirected graph.
4. He used BIESO based on Natural Languages Processing is used to recognize the word from sentence.

Disadvantages

1. His proposed model does not work on table which have no layout (rows and column line).

Schreiber, et al. [3], proposed a deep learning-based approach to detect table and structure recognition of table in Document or Image by using the ICDAR 2013 table completion dataset which containing 67 documents with 238 pages. Author proposed two architecture based on deep learning solution -

1. Table Detection

In this article the author proposed architecture, the author used F-RCNN based framework for table detection. F-RCNN framework is divided into two parts: First, RPN produces a region proposal depend on the image input. Afterward, these proposals are classified by the Faster-RCNN network. The author also used a ZFNet and weight of the VGG-16 network. In the author's experiments, he also proposed a thought about the Marmot C dataset that does not contains more images for training the deep neural network from scratch. In this proposed model, the author using raw images instead of processing the PDF documents contains metadata which makes it more challenging.

2. Table Structure Recognition

After successful detection of the table in Document or Image. The most epochal challenge for the author is to identify the location of rows and columns to extract information from it by implementing the full connected network with the weight of VGG-16.

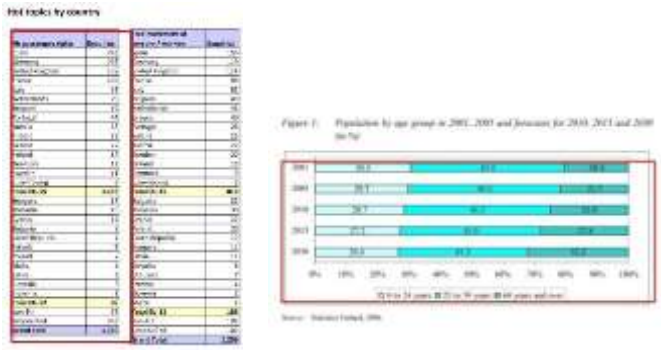
Advantages

1. It has a feature of table structure recognition while [4] proposed model do not have.
2. In His proposed model, the author also shows some successful table detection result (Figure1).

3. His model can also detect row and column to make a physical structure of table.
4. His proposed model is also capable to clean out the ground-truth annotations of dataset.
5. His model used only raw images without no additional data to make more challenging.
6. He increased the amount of background separation in pre-processing model from the remaining structure components.
7. His proposed model can also detect the multiple table, large table, small table and page column.
8. His model performance evaluation result on recall and precision for table detection (96.15% and 97.40%) and for structure recognition (87.36% and 95.93%) which is far better than above other discussed model.

(a) (b)

Figure 1: Table Detection Result on ICDAR 2013 dataset [3]



(a) (b)

Figure 2: Some failure result [3]

Disadvantages

1. His model also fails when images having close columns, nested row hierarchy and very close by row (Figure 2).
2. Has a bar chart confusion.
3. His model has a persisting issues with recognizing structure in images which has very nearest proximity to other elements.

Vasileiadis, et al. [12]**,** proposed a method to extract tabular data from document image by using OCR and following the set of heuristic-based rule in order to detect tabular structure, based on the imagination that text table are areas of strictly formatted text. The document images are exported to the simple HTML format if the table structure present in the documents. The author proposed his architecture in two models –

1. Image Processing

In initial steps, binary image is changed size and resampled in order to approximate a 300 dpi (Dots per inch) A4 Size paper and applied tesseract OCR to get the editable text from document image. The author fed the raw text to line detector to delete any visible grid lines and also discard the non-text areas as (Figure 1.a.). The text-only image is looping searched for continues horizontal empty space, which helps to indicate the multi-column areas. The algorithm select subspace between horizontal (left to right) empty areas and within them for Vertical (top to bottom) empty space through the whole height of subspace. If any area found that are considered as potential document column separators, and if the standard deviation of their width is smaller than a threshold□□h□□□h, the subspace is registered as multicolumn text area. All the detected multi-column text areas are sequentially vertically reordered, in top-to-bottom, left-to-right order.

2. Table Reconstructions

In this proposed second method, for table detection and table reconstruction the author includes four steps, each following a set of heuristic-method in order to recognize the table structure.

2.1 Text Lines and Word Segmentation

The words make a text lines overlaps vertically, using the certain coordination of each words. The word segmentations are defined by converging words together for each lines. The author defined a three types of line segmentation. 1) Text – Line with a single expanded segmentation. 2) Table – Lines contain more than one segmentation. 3) Unknown – which does not follow any above types of segmentation. The author proposed if the document having a multiple page, top and bottom areas also be checked for same repetitive word segment.

2.2 Initial Table Areas

The table areas detect if having a more than one line segment.

2.3 Table Column Generation

The author proposed an algorithm for both single column segment and multi-column segment.

2.4 Multiple-lines Table Rows

Some of table rows integrate into multiple-line rows.

Advantages

1. His proposed model can do the multiple column and single column segmentation.
2. His proposed is based on the heuristic method for table detection and extraction from table.
3. It is discard the non-text areas.
4. All the detected multicolumn text areas are sequentially
   vertically reordered, in top-to-bottom, left-to-right order.
5. His proposed method achieved high precision and recall rates: 88.30% and 97.22% for tables, and 89.45% and 93.70% for cells respectively.

Disadvantages

1. Cell accuracy can be decreased if the table data will not be well aligned.
2. His proposed model's algorithm efficiency is not good.
3. His proposed method also do not work with table do not contains any row and columns line separator.

3. CONCLUSIONS

The main objective of this paper is to show best method based on my survey. There are two best method (Schreiber, et al., 2017**)** is proposed in model based on F-RCNN algorithm and also using RPN to identify region with his model. Based on his proposed model he is detecting table along with structure recognition but it also have some limitation which we do have in (Vasileiadis, et al., 2017) proposed his model in two parts 1) image preprocessing – where he used it to make image accurate to process in next step after applying some filtration. 2) In this step he proposed better approach to detect table and table reconstruction by applying heuristic-method.

ACKNOWLEDGEMENT

I would like to express my profound gratitude to professor Dr. MN Nachappa and Prof Subarna Panda, for their patient, encouragement and valuable assessments of this research work. I appreciate his willingness to generously contribute time.

REFERENCES

[1] A. S. Tarawneh, A. B. Hassanat, D. Chetverikov, I. Lendak and C. Verma, "Invoice classification using deep features and machine learning technique," *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT),* pp. 855-859, 2019.

[2] S. Siddiqui, M. Malik, S. Agne, A. Dengel and S. Ahmed, "DeCNT: deep deformable cnn for table detection," *IEEE,* vol. 6, p. 12, 2018.

[3] S. Schreiber, S. Agne, I. Wolf, A. Dengel and S. Ahmed, "DeepDeSRT: deep learning for detection and structure recognition of tables in document images," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),* p. 6, 2017.

[4] A. Gilani, S. Qasim, I. Malik and F. Shafait, "Table detection using deep learning," *IEEE,* p. 6, 2017.

[5] J. Fang, P. Mitra, Z. Tang and C. Giles, "Table header detection and classification," *AAAI'12: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence,* pp. 599-605, 2012.

[6] G. Naganjaneyulu, N. V. Sathwik and A. Narasimhadhan, "A multi clue heuristic based algorithm for table detection," *2016 IEEE Region 10 Conference (TENCON),* vol. 1, 2016.

[7] S. Buzykanov;, "Enhancement of poor resolution text images in the weight sololev space," in *2012 19th International Conference on Systems*, Vienna, Austria, April 2012.

[8] T. Dhiran and R. Sharma, "Table detection and extraction from image document," *International Journal of Computer & Organisation Trends,* vol. 3, no. 4, 2013.

[9] F. Shafait and R. Smith, "Table detection in heterogeneous documents," *The Ninth IAPR International Workshop on Document Analysis Systems,* pp. 1-8, 2010.

[10] T. Huynh-Van, T. L. B. Khanh and T. A. Tran, "Information learning to detect tables in document images using line and text," *ICMLSC '18: Proceedings*

of the 2nd International Conference on Machine Learning and Soft Computing, p. 151–155, 2018.

[11] F. Kleber, H. Dejean and E. Lang, "Matching table structures of historical register books using association graphs," pp. 217-222, 2018.

[12] M. Vasileiadis, N. Kaklanis, K. Votis and D. Tzovaras, "Extraction of tabular data from document images," *W4A '17: Proceedings of the 14th Web for All Conference on The Future of Accessible Work,* p. 2, 2017.