Table Detection and Extraction from Image Document

Tanushree Dhiran[#1], Rakesh Sharma[*2]
*M.Tech Scholar[#1], Assistant Professor[*2]*
RCEW, Jaipur, India

***Abstract*—** *Tables make information easier to understand and perceive than regular text block. Now days, it becomes popular structure for information representation. Format of tables differs and change according to need of representation of information. Various format of table makes it difficult for OCR system to recognize and just segment as an Image block. We proposed a novel approach which can detect all type of table format from single column image document. Tables are categorized in three type based of their rows and column separator.Type1 table have line as row and column separator. Type2 table have horizontal line for separating rows and space for separating column. In Type3 tables only space are used as both row and columns separator. Tables are detected from image documents based on simple projection profile and hough line detection method. We have tested this approach with 1200 image documents which contains all type of table format and get 89% accurate result.*

***Keywords*—** Line Segmentation, Hough Line Detection, Word Level segmentation, Projection Profile

I. INTRODUCTION

Table is an efficient way used for representation of information. Now days Tables are presents in almost all type of documents like in magazines, news papers, books, etc...Table detection becomes important because it needs different layout analysis. Different structures of table makes difficult for OCR engine to detect the table. There are some works already done for this purpose. Zheng, Liu,Ding and Pan in 2001 detect frame line based on DSCC method. In this paper, Line gets detected with determining the black run length and can be merged with other line if certain condition gets satisfied. This approach is not work with table in which wide spaces are used as row and column separated because it depends on Line detection [1]. Kieninger proposes an algorithm which is based on robust block segmentation. It takes word bounding box as input, make a cluster of words. Now, determine that it is a table region or not based on some heuristics and rule [2]. Shafait and Smith in 2010 work on multi-column document. They first determine the column gap between columns of page and then determining the table based on horizontal ruling [3]. Klein proposed a method for Industrial document analysis in 2010. There are three steps in his proposed approach which includes: table header search based on known set of headers, searching for tabular structure and searching for groups of line [4]. Zuyev introduced a concept of table grid. Based on some hypothesis and classification rule with threshold value Table can be detected [5]. These above described approaches are not work with all type of Tables structure. We have categorized tables into 3 parts as shown in figure.

Amit	10	Delhi
Sunil	12	Mumbai
Ajay	15	Pune
Rakesh	16	Delhi
Pankaj	12	Jaipur

Figure 1: Table of Type1 which has line as row and column separator.

Item Name	Width	Length
Rectangle	4	12
Square	4	4
Rhombus	4	5

Figure 2: Type2 tables where space is used to separate Columns and line for rows.

Susan	492	November	Completed
Bill	625	December	Pending
George	333	May	Pending
John	129	July	Cancelled

Figure 3: Type3 table where space is used to Separate columns.

II. PROPOSED SCHEME

Pre-processing is the initial and essential part of any OCR processing. It involves binarization, noisy border removal and enhancement. Binarization is the process of converting colour image or gray scale image to binary image. It is accomplished taking some appropriate value as threshold value. Pixels having greater intensity than threshold value changed into black otherwise white intensity value. Gatos has given an adaptive method of binarization [6]. Perantonis given an approach for skew correction based on

Hough transformation and rectangular bounding box. Noisy border removal can be done using flood fill algorithm [7]. For enhancement, dilation process is applied. In this process a suitable structuring element is taken so that small gaps which occur as a noise must be removed. Structuring elements in rectangular shape are most suitable for joining lines.

White space between text lines is used to segment the text lines. The line segmentation is carried out by calculating the horizontal projection profile of the whole document. The horizontal projection profile is the histogram of number of black pixels along every row of the image. The projection profile exhibits valleys of zero height corresponding to white space between the text lines. Line segmentation is done at these points. Similarly White space between text words is used to segment the text lines. Word segmentation is done by the vertical projection profile. The vertical projection profile is the histogram of number of black pixels along every column of the segmented line. The projection profile exhibits valleys of zero height corresponding to white space between the text words.

We are using combination of two methods to detect Tables. Method1 is used for detecting Type1 tables where as Method2 is used to detect Type2 and Type3 tables. At first we define some features and constraints that should be satisfied by tables to be detected, such as:

- A document or page can have four types of blocks: (1) text block(2) image block (3) table block
  (4) blank block.
- Columns are separated by more than word space (for Type1 table)
- Columns are aligned (almost) (for Type1 table).
- Minimum dimension of Table must be 2*2.
- There is no effect of font size or style, if all texts in that document use same size and style.

Steps of Method for Type1 table detection:

1. Calculating the average height of a text line based on horizontal projection on page image.
2. All the connected components which has height less than twice of text line height get removed. We have taken twice height because minimum dimension of table is 2*2.
3. Now, only image or tables may be present in page document.
4. In image, intensity value from one pixel to another pixel changes slowly in comparison to Tables. So now image block get also removed.
5. With the help of Hough transformation lines get detected. We also calculate the intersection points of lines. Based on these two structure of table get extracted.

Figure 2 showing the feature of the table of type1 which is extracted and the figure3 showing the output of the system.

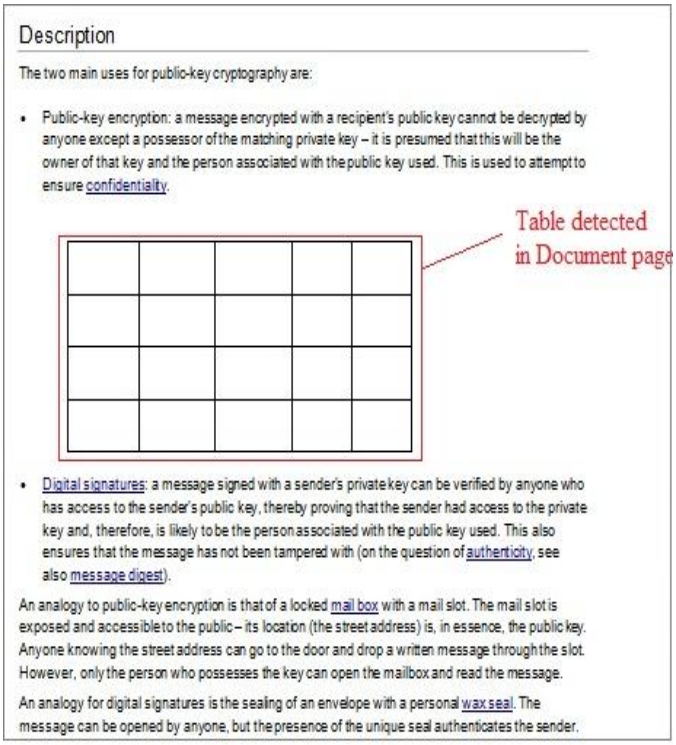

Figure 4: Type1 table Detection in Document Image

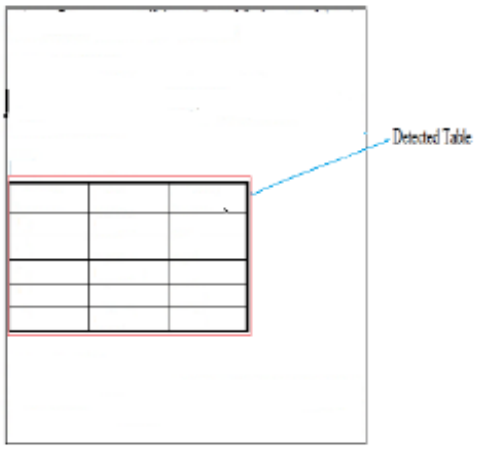

Figure 5: Type1 Table detected and extracted

Steps of Method2 for Type2 and Type3 table detection:

1. With the help of horizontal projection text line segmentation has done.

2. With the help of Hough transformation, horizontal lines get detected and removed if any found in page Image.

3. We choose one text line where number of words is maximum and taken largest word space from that line as threshold value.

4. Takes one text lines as input, with the help of vertical projection on lines if there is any space found which length is significantly larger than calculated word space (threshold value) then these lines are part of a table.

5. Takes next text line as input and combine it with previously taken input text lines.

Steps 4 and 5 get repeated until we found that current input text line is also part of table. Otherwise next Input line is not combined with previously taken lines. Again go to step4 up to last text line.

III. EXPERIMENTAL RESULTS

The above described approach is applied to 1200 numbers of input images. These images are scanned pages from books, forms, magazines, certificates etc. In this collection of testing images 450 page documents have Type1 tables, 600 image documents have Type2 or Type3 tables of different layouts, 150 image documents do not have any type of tables. This approach for detection and extraction is implemented in Visual C++ using OPENCV. Input images are tested over the system in both one by one and also in batch processing. The graph below shown here for the testing result of the system as it gives almost 89 % accuracy over 1200 image documents of different pattern.

TABLE I
TESTING RESULTS

Category of tables	Total number of images	Total number of tables present in these images	Correctly detected number of tables	Wrongly Detected Tables
Type1	450	542	454	88
Type2 & Type3	600	722	668	54
No Table	150	0	-	0

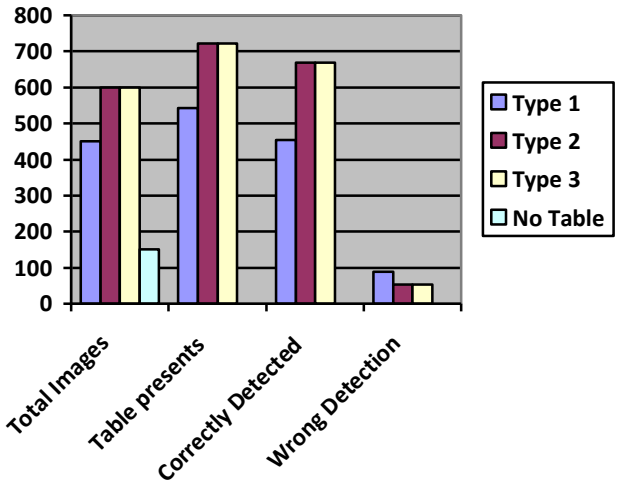

Figure 5: Graph Showing Table detection Results

Result graph depicted here showing percentage of correct detection and wrong detection of the table present in document image. In some cases a connected component which behaves like a table detected with this system. No table level of the graph showing number of table detected which is in fact not a table. This type of wrong detection can be solved in the next publication of the authors.

IV. CONCLUSIONS

The above method is applied for automatic detection of Table Detection from document images. Tables are categorized in three type based of their rows and column separator.Type1 table have line as row and column separator. Type2 table have horizontal line for separating rows and space for separating column. In Type3 tables only space are used as both row and columns separator. Experimental results demonstrate the efficiency of the proposed method. This approach can be extended with the multi column page documents. For the experiments of this method we use visual C++ with OPEN CV as a tool.

ACKNOWLEDGMENT

The author is a RTU Research Fellow and the work presented here was done while M Tech Research at RCEW College, Jaipur. The author is grateful to Dr. C S Lamba and Ms. Vandna Verma. Authors are thankful to Mr. Vinit Khanna for his conversations and suggestions.

REFERENCES

[1] Zheng Y., Liu C.(2001), Ding X., Pan S.,"Form Frame Line Detection with Directional Single-Connected Chain", Proc. of the 6th Int. Conf. on Doc. Anal. & Recognition, 699-703

[2] Thomas G. Kieninger(1998), "Table Structure Recognition Based on Robust Block Segmentation", German research center for artificial intelligence

[3] Shafait and Smith (2010), "Table Detection in Heterogeneous Documents", DAS '10 Proceedings of the 9th IAPR International Workshop on Document Analysis Systems.

[4] B. Klein, S. Gokkus, T. Kieninger, A. Dengel(2001), "Three approaches to "industrial" table spotting", Sixth International Conference on Document Analysis and Recognition (ICDAR01), Seattle, WA, September, pp.513–517.

[5] K. Zuyev(1997), "Table image segmentation", Proceedings of the International conference on Document Analysis and Recognition (ICDAR) '97, Ulm, Germany, August, pp.705–708.

[6] Gatos, B., Pratikakis, I., Perantonis, S.J(2004): "An adaptive binarisation technique for low quality historical documents". IARP Workshop on Document Analysis Systems (DAS2004), Lecture Notes in Computer Science (3163), ,pp.102-113

[7] Perantonis, S.J., Gatos, B., Papamarkos, N(1999) "Block decomposition and segmentation for fast Hough transform evaluation". Pattern Recognition, vol. 32(5), pp.811-824

[8] B. Gatos, D. Danatsas, I. Pratikakis and S. J. Perantonis. Automatic(2005) ,"Table Detection in document images". National Center for Scientific Research "Demokritos"GR 15310 Athens, Gre