Detection of Table Structure and Content Extraction From Scanned Documents

S.Deivalakshmi, K.Chaitanya and P.Palanisamy

Abstract— **Tables are one of the efficient information conveying methods used now days in larger extent. This paper report a fast, language independent (English and Tamil), skilled technique for table structure detection and its content extraction from a scanned document image based on morphological operation, connected components and labeling. From the conducted exhaustive experimentation, it is observed that the proposed method is the fastest approach because of its simple operations. In addition with that it is noticed that it does not lead to any kind of degradation in the extracted table content since after detecting contents location it is retrieved from the original image. More over it is also very interesting to note that the presented approach works well for documents with different font's size and font styles.**

***Index Terms*—scanned document image; table detection and content extraction; morphological operation; connected components, labeling;**

I. INTRODUCTION

DUE to the rapid growth in the technology the information is increasing day-by-day. The world's information double's every two years [1], to store and use/ process this information we need huge digital library space for sorting. But with the aid of pictorial representation in most of the cases huge information can be conveyed with less space.

The numerous breakthroughs in the emerging area of digital image processing (DIP) have extended its grasps to the fields of remote sensing, satellite imaging, biomedical imaging, document imaging, astronomy, geology etc. One of the emerging field of DIP is document image processing. The traditional source of storing data is paper. But in contemporary circumstances, all documents from modern office agreements to valuable ancient records are being digitized for processing and storage purposes. Document image analysis can be segmented into two parts: Text processing and Graphical processing. Text processing deals with recognition of text, words [2], textual lines, skew [3, 4] etc. Graphical processing deals with non-textual elements such as tables, lines, images, symbols, etc. Tables are one of the pictorial representation techniques seen in all type of documents such as newspapers, magazines, books, etc. There is no unique style/ format/layout for tables which make the OCR engine difficult to recognize. Table detection, segmentation and extraction is not a new work [5, 6, 7, 8, 9, 10, 11], there are already some techniques proposed which will be described as follows.

The detection of Frame Line based on DSCC method was put forth by Zheng et.al. in 2001. In their work they proposed a technique to detect lines [12] and merge it with other lines based on certain conditions. However this approach does not work for table with wide spaces which separate rows and columns. Robust Block segmentation technique was proposed by Kieninger et.al. in the form of an algorithm. In this technique each individual box around the information content of a table are taken as input and are put together. The structure so formed is recognized to be a table or not based on some heuristics and rules [13]. A method was proposed by Laurentini and Viada [14] to detect tables based on horizontal and vertical projection profiles in which the text blocks in an area are compared with the pattern of lines detected. A top-down approach for table analysis was put forth by Green and Krishnamoorthy [15] in which the features of table region are obtained from Horizontal lines, vertical lines, horizontal space and vertical space. Elementary cell characterization is performed to label individual cells and the labels are matched to a table model such that the relational information can be extracted.

Automatic table detection method proposed by Gatos et.al. [16] used nine different types of masks for detecting lines intersection in a table. S. Mandal et.al proposed a simple and effective system for table detection from document image but in this method more than 48% of the page should contain table [17]. Work on multicolumn documents was carried by shafait et.al. in the year 2010. They proposed a scheme to find the gap between columns of a page and then determine table based on horizontal ruling [18]. A Method for analysis of industrial documents was proposed by Klein et.al. in the year 2010. The three steps put forth by them were: To search for table header based on Knowledge of headers available, to search for structure of table and search for groups of lines [19]. The concept of Table grid was introduced by Zuyev et.al. According to his proposal Table can be detected based on few classification rules with threshold value [20]. System proposed by Tanushree Dhiran et.al. for Table Detection and Extraction from Document image has its drawback as it should contain

S.Deivalakshmi.is with the Department of Electronics and Communication Engineering, National Institute of Technology, Trichy- 620015, phone: 0431-2503321; (e-mail: deiva@nitt.edu) .

K.Chaitanya , is with the Department of Electronics and Communication Engineering, National Institute of Technology, Trichy- 620015 (e-mail: seehaitanya@gmail.com).

P.Palanisamy is with the Department of Electronics and Communication Engineering, National Institute of Technology, Trichy- 620015, phone: 0431-2503312; (e-mail: palan@nitt.edu).

978-1-4799-3358-7/14/$31.00 ©2014 IEEE

same font size throughout the document which in reality is not the case [21].

This paper is organized as follows. Proposed work is dealt in Section II. Experimental results and discussions are presented in Section III. Finally concluding remarks are given in Section IV.

II. PROPOSED METHOD

The proposed approach aims to present a fast, language independent (English, and Tamil) table structure detection and its content extraction from a scanned document image. This method consists of two steps:

(A) Detection of table structure using morphological operation.

(B) Extraction of detected table contents based on connected components and labeling.

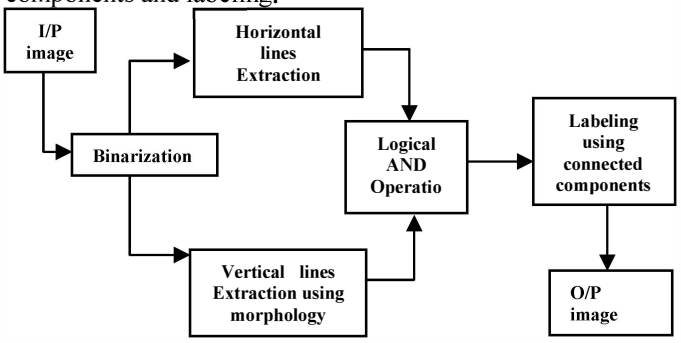

Fig.5 System Block diagram for Table Contents extraction

The first and foremost step is pre-processing in any OCR. It involves binarization, noisy border removal and enhancement. Binarization is the process of converting colour image or gray scale image to binary image. As given in Fig.5, input image is binarized (binarization) then horizontal & vertical line extraction module is performed on the binarized input. Carrying out logical AND operation on extracted vertical and horizontal lines gives table structure. Here scanned document in binary form is considered as an input as shown in Fig.1 (a).

16. If the number of students passing an examination be considered a criteria for comparison of difficulty level of two examinations, which of the following statements is true in this context ?
(a) Half-yearly examinations were more difficult.
(b) Annual examinations were more difficult.
(c) Both the examinations had almost the same difficulty level
(d) The two examinations cannot be compared for difficulty level.
(e) For students of Sections A and B, the annual examinations seem to be more difficult as compared to the half-yearly examinations.

Directions (Questions 17 to 21) : *The following table shows the number of new employees added to different categories of employees in a Company and also the number of employees from these categories who left the company every year since the foundation of the Company in 1995.* (Bank P.O. 2001)

Year	Managers		Technicians		Operators		Accountants		Peons	
	New	Left	New	Left	New	Left	New	Left	New	Left
1995	760	—	1200	—	880	—	1160	—	820	—
1996	280	120	272	120	256	104	200	100	184	96
1997	179	92	240	128	240	120	224	104	152	88
1998	148	88	236	96	208	100	248	96	196	80
1999	160	72	256	100	192	112	272	88	224	120
2000	193	96	288	112	248	144	260	92	200	104

Fig.1(a) Scanned document image in binary form


ராவில் ரன் அவுட் ...ர்கி வெளி யே தினார். சன்ரை சர்ஸ் 2 ஓவரில் 12 ரன்னுக்கு 2 விக் கெ. இழந்து தடு மாறியது. இந்த லையில் தவான் கேப்டன் கேமரான் ட் ஜோடி பொறுப் ற விளையாடி. ரன் ர் க ர் து, ரூபாயி ரடியில் தவான். மோகித் பதம் பவனை சன்ரை கணிச து. இந்த ஓவரில் 2 எடுத்து ஆத்தில் ர் டோனி ட்டது டைவை அபித் ள் சே மாற்றத் ாயாடிய ன் சுழ

வ்கோர் விவரம்		
டி காக்	(தி) விஜய் (பி) மோகித்	4
தவான்	(கு.ப.பிழக்கவில்லை)	63
ஹஷிம ஆம்லா	(ரன் அவுட்)	2
கேமரான் ஒயிட்	(தி) டோனி (பி) மோகித்	2
மில்லர்	(தி) பிராவோ (பி) தவ்வின்	15
சற்குணம்	(தி. பி) பிராவோ	10
டேன் சம்மி	(தி) மோரிஸ் (பி) பிராவோ	19
ஆசிஷ் ரெட்டி	(கு.ப.பிழக்கவில்லை)	36
உதிரிகள்		8
மொத்தம்	(6 விக்கெட், 20 ஓவர்)	159
விக்கெட் வீழ்ச்சி: 1-4, 2-12, 3-37, 4-61, 4-83* (தவான் காயம்), 5-98, 6-116.		
சென்னை பந்துவீச்சு: மோகித் ஷர்மா 4-0-33-2, ஜேசன் ஹோல்டர் 4-0-31-0, அஸ்வின் 4-0-21-1, மோரிஸ் 3-0-24-0, பிராவோ 4-0-37-2, ஜடேஜா 1-0-8-0.		

விடிய்டார். அடுத்து அரீமுக வீர் தலைவன் சற்குணம் களமிறு கினர். பத்து ஓவரில் சன்ரைசர்ஸ் 4 வீக்கெட் இழப்புக்கு 62 ர எடுத்தது. அரை சதம் அடிப்பா னை எதிர்பார்க்கப்ப நிலையில் தவான் 44 ரன் எடுத்து (38 பந்து 6 பவுண்டரி) கடாயம் காரணமா வெளியேன் திரும்பினார். ச குணம் 10 ரன் எடுத்து பிராவே வேகத்தில் அவரிடமே விடிய்டா சம்மி 19 ரன்னில் வெளியேறினா இதைத் தொடர்ந்து, மீண்டு உள்ளே வந்த தவான் அடுத் டுத்த பவுண்டரி விளாசி அரைசத கட்டார்.

கடைசி கட்டத்தில் தவான் ஆசிஷ் ஜோடி அதிரடியில் இறங் ஸ்கோர் கணிசமாக எகிறிய பிராவோ வீசிய கடைசி ஓவரில் ரன் கிடைத்தது. சன்ரைசர்ஸ் ஓவரில் 6 விக்கெட் இழப்புக்கு 1 ரன் குவித்தது. தவான் 63 ரன் (பந்து, 10 பவுண்டரி), ஆசிஷ் ரெட் 36 ரன் (16 பந்து, 2 பவுண்டரி, சிக்சர் எடுத்து ஆட்டமிழக்காம இருந்தனர்.

சூப்பர் கிங்ஸ் பந்துவீச்சி மோகித், பிராவோ தலா 2, அ வின் ஒரு விக்கெட் வீழ்த்தின அடுத்து 20 ஓவரில் 160 ர எடுத்தால் வெற்றி என்ற சற்ற கடினமான இலக்குடன் சென்ன களமிறங்கியது.

Fig.2(a) Scanned document image in binary form

A. Detection of table structure using morphological operation

A binary document image may have four types of blocks:

(i) Text block (ii) Image block (iii) Table block and (iv) Blank block.

Of these blocks, detection of table is based on the observation that a table contains intersection of horizontal and vertical lines. Lines are formed by group of running black pixels. The horizontal and vertical lines are detected by taking two different structuring elements such as Horizontal structuring element (H) and vertical structuring element (V). Based on experimentation it is found that dilation alone can detect vertical and horizontal lines but precise results can be obtained using closing operation. For detection of horizontal lines we used the mathematical morphological operation as given in equation (1).

$$J = ((I \oplus H)\Theta H) \tag{1}$$

Similarly for detection of vertical lines we used the mathematical morphological operation as given in equation (2).

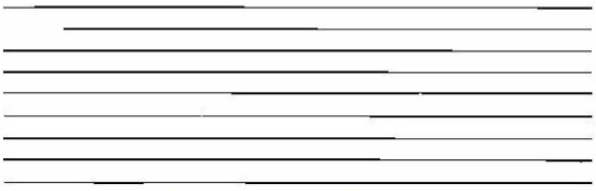

Fig.1(b) Extracted Horizontal lines of Fig.1(a)

$$K = ((K \oplus V)\Theta V) \tag{2}$$

Here J as shown in Fig.1 (b) & K as shown in Fig.1 (c) are extracted horizontal and vertical lines of considered input image(I).

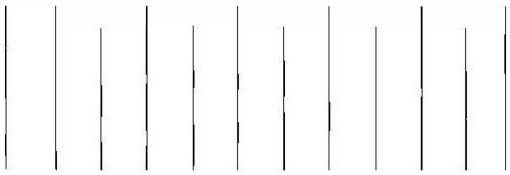

Fig.1(c) Extracted Vertical lines of Fig.1(a)

$$L(m,n) = J(m,n)ANDK(m,n) \quad (3)$$

Next as given in equation (3) Logical AND operation is performed between the images of extracted horizontal and vertical lines of considered input image(I) which in turn gives us table structures. Finally the image L as shown in Figs.1 (d), 2(b) contains only the structure of tables.

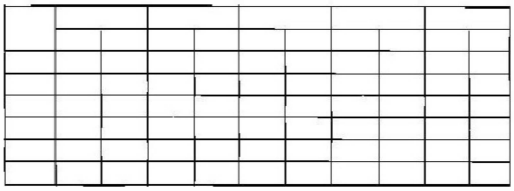

Fig.1(d) Extracted Table structure of Fig.1(a)

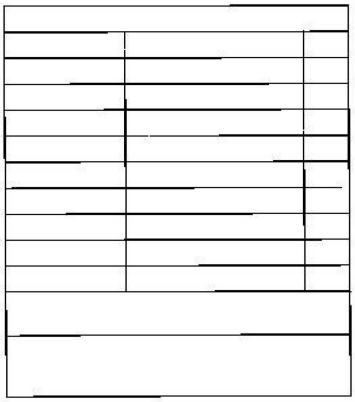

Fig.2(b) Extracted table structure of Fig.2(a)

B. Extraction of table contents using connected components and labelling

The image containing table structure (L) is taken as input for extracting table contents. The image (L) contains only the table structure and blank space, which is labeled based on 4 connected components [22].

For example the matrix given in Fig.4 (a) is pixel representation of image (L). Here all 1's indicate blank space within and out off table structure. The resultant labeled output for this is matrix given in Fig.4 (b).

1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	0	0	0	0	0	0	0	1	1
1	1	0	1	1	0	1	1	0	1	1
1	1	0	1	1	0	1	1	0	1	1
1	1	0	0	0	0	0	0	0	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1

Fig.4 (a) Matrix Representation of image L

By observation it is noticed that labels containing 1's refer to text content and blank space outside the table structure in input image. Coordinates of 1's are collected from image(L) and It's corresponding information from original image(I) are stored in an array, this in turn gives us text content outside the table structure. This extracted image is a table free image (E) as shown in Fig.1 (e).

16. If the number of students passing an examination be considered a criteria for comparison of difficulty level of two examinations, which of the following statements is true in this context ?
(a) Half-yearly examinations were more difficult.
(b) Annual examinations were more difficult.
(c) Both the examinations had almost the same difficulty level.
(d) The two examinations cannot be compared for difficulty level.
(e) For students of Sections A and B, the annual examinations seem to be more difficult as compared to the half-yearly examinations.

Directions (Questions 17 to 21) : ***The following table shows the number of new employees added to different categories of employees in a Company and also the number of employees from these categories who left the company every year since the foundation of the Company in 1995.*** **(Bank P.O. 2001)**

Fig.1(e) Table free image of Fig.1(a)

As given in equation (4) by subtracting the table free image (E) from the original image (I) gives us desired image (T), where the image (T) as shown in Fig.1 (f) contains only table structure along with the table contents which is our desired region of interest.

$$T(m,n) = I(m,n) - E(m,n) \quad (4)$$

Year	Managers		Technicians		Operators		Accountants		Peons	
	New	Left	New	Left	New	Left	New	Left	New	Left
1995	760	—	1200	—	880	—	1160	—	820	—
1996	280	120	272	120	256	104	200	100	184	96
1997	179	92	240	128	240	120	224	104	152	88
1998	148	88	236	96	208	100	248	96	196	80
1999	160	72	256	100	192	112	272	88	224	120
2000	193	96	288	112	248	144	260	92	200	104

Fig.1(f) Extracted table with content of Fig.1(a)

1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	0	0	0	0	0	0	0	1	1
1	1	0	2	2	0	3	3	0	1	1
1	1	0	2	2	0	3	3	0	1	1
1	1	0	0	0	0	0	0	0	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1

Fig.4 (b) Matrix Representation of Labeled output.

III. EXPERIMENTAL RESULTS

The proposed approach for table detection and its content retrieval is implemented using Matlab. For detecting horizontal lines we used a straight line structuring element of size 1×35 with all ones as its elements. Similarly, for detecting vertical lines we used a straight line structuring element of size 35×1 with all ones as its elements. We applied this method not only to unique table structures in different languages such as English and Tamil but also to tables containing graphics. The Proposed method is applied on 150 images taken from scanned newspapers, textbooks and tobacco800 database images, which contains different table structures. This method can extract any form of table contents even if it is not text information. There is no restriction on percentage of table content in this method as in [17]. Gatos [16] used morphological operation as given in equation (5) to detect horizontal lines.

$$IM_H = (IM \bigcup (((IM \Theta B_{HR}) \bigcup (IM \Theta B_{HL}) \oplus B_H)) \quad (5)$$

Similar equation was used for vertical line detection also.

Draw back of his method is that, if there is a small brake in the lines then mask cannot detect that portion of table. So, to overcome this, he first enhanced the line brakes using above equation. But the proposed method works well without enhancing small brakes introduced by binarization. From the considered 150 images for experimentation, there were 1455 horizontal lines, 962 vertical lines and 251 tables. Out of which 1397 horizontal lines, 945 vertical lines and 229 tables were detected successfully by this proposed method as given in Table.1. It is observed that because of the binarization process some of the lines are missing or degraded in few images that is why it is not possible to detect all the lines effectively by this method. For images having no lose of lines or degradation of lines due to binarization this method gives 100% throughput. By considering all the effects due to binarization this technique gave 91% throughput. From the conducted exhaustive experimentation, it is observed that the proposed method is the fastest approach because of its simple operations. In addition with that it is noticed that it does not lead to any kind of degradation in the extracted table content since after detecting contents location, content or information is extracted from the original image. More over it is also very interesting to note that the presented approach works well for documents with different font's size and font styles.

Parameters	Present	Detected and extracted
Horizontal line	1455	1397
Vertical lines	962	945
Table	251	229

Table.1 Experimental results for Table detection

IV. CONCLUSION

From the conducted exhaustive experimentation, it is observed that the proposed method is the fastest approach because of its simple operations. More over the proposed approach in this paper is font and language independent. It is important to note that this method has two limitations: i) because of the binarization process some of the lines are missing or degraded in few images that is why it is not possible to detect all the lines effectively by this method. This can be overcome by means an effective binarization technique which in turn leads to better results than what we obtained now. ii) This method is not applicable if there are no lines on the input document or documents with tables made up of only one kind of lines i.e, horizontal/vertical lines alone. Above mentioned two drawbacks provides future scope for the improvement of the proposed method.

REFERENCES

[1] http://www.emc.com/leadership/programs/digital-universe.htm

[2] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros and George Carayannis,"Handwritten document image segmentation into text lines and words" Pattern Recognition Vol. 43, pp. 369–377, 2010.

[3] Rosner, D., Boiangiu, C-A., Stefanescu, A., Tapus, N. and Olteanu, A., "Text line processing for high-confidence skew detection in image documents", IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2010) , pp.129–132, 2010.

[4] Y.Y. Tang, S.W. Lee and C.Y. Suen, "Automatic document processing: A Survey", Pattern Recognition Vol. 29, Issue 12, pp. 1931–1952, 1996.

[5] Chandran, S., Balasubramanian, S., Gandhi, T., Prasad, A.Kasturi, R., Chhabra, A. "Structure recognition and information extraction from tabular documents", IJIST, pp. 289–303, 1996.

[6] Das, A.K., Chanda, B. "Detection of tables and headings from document image: a morphological approach", International Conference on Computational linguistics, Speech and Document Processing (ICCLSDP'98), pp. A57–A64, 1996.

[7] Tsuruoka, S., Takao, K., Tanaka, T., Yoshikawa, T., Shinogi, T. "Region segmentation for table image with unknown complex structure", in proceedings of ICDAR'2001, pp. 709–713, 2001.

[8] Watanabe, T., Luo, Q.L., Sugie, N. " Layout recognition of multikinds of table-form documents", IEEE Trans. on Pattern Anal.Machine Intell. , pp. 432–446, 1995.

Hu, J., Kashi, R., Lopresti, D., Wilfong, G." Medium-independent table detection", in SPIE Document Recognition and Retrieval, pp. 291–302, 2000.

[9] Ramel, J.-Y., Crucianu, M., Vincent, N., Faure, C. "Detection, extraction and representation of tables", in the procedings of 7th International Conference on Document Analysis and Recognition, vol. 1, pp. 374–378, 2003.

[10] Tersteegen, W.T., Wenzel, C." Scantab: table recognition by reference Tables". in Proceedings of 3rd IAPR workshop on Document.

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

[12] Thomas G. Kieninger, "Table Structure Recognition Based on Robust Block Segmentation", German research center for artificial Intelligence, 1998.

[13] A. Laurentini and P. Viada, "Identifying and understanding tabular material in compound documents", Proc. Intl. Conf. Patt. Recog., 1992.

[14] E. A. Green and M. S. Krishnamoorthy," Model-Based Analysis of Printed Tables", Proc. Intl. Conf. Doc. Anal. and Recog., pp. 214- 217, 1995.

[15] B. Gatos, D. Danatsas, I. Pratikakis and S. J. Perantonis, "Automatic Table Detection in Document Images". Proceedings of the third international conference on Advances in Pattern Recognition (ICAPR'05) Vol. 1, pp. 609 – 618, 2005.

[16] S. Mandal, S. P. Chowdhury ,A. K. Das and Bhabatosh Chanda "A simple and effective table detection system from document images", International Journal of Document Analysis, pp.172–182, 2006.

[17] Shafait and Smith ,"Table Detection in Heterogeneous Documents", Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp.65-72, 2010.

[18] B. Klein, S. Gokkus, T. Kieninger, A. Dengel, "Three approaches to "industrial" table spotting", Sixth International Conference on Document Analysis and Recognition (ICDAR01), pp.513–517, Sep2001.

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

[20] Tanushree Dhiran and Rakesh Sharma,"Table Detection and Extraction from Image Document", International Journal of Computer & Organization Trends, pp275-277, 2013.

[21] Haralick, Robert M., and Linda G. Shapiro, Computer and Robot Vision, Volume I, Addison-Wesley, pp. 28-48, 1992.