https://dx.doi.org/10.22161/ijaers.4.5.13 *ISSN: 2349-6495(P) | 2456-1908(O)*

Segmentation of Unstructured Newspaper Documents

Santosh Naik[1], R. Dinesh[2], Prabhanjan S.[3]

[1]Research Scholar, Department of Computer Science and Engineering, School of Engineering and Technology, Jain University, Bangalore, India
[2]Research Supervisor, Department of Computer Science and Engineering School of Engineering and Technology, Jain University, Bangalore, India
[3]Research Scholar, Department of Computer Science and Engineering, School of Engineering and Technology, Jain University, Bangalore, India

Abstract— *Document layout analysis is one of the important steps in automated document recognition systems. In Document layout analysis, meaningful information is retrieved from document images by identifying, categorizing and labeling the semantics of text blocks from the document images. In this paper, we present simple top-down approach for document page segmentation. We have tested the proposed method on unstructured documents like newspaper which is having complex structures having no fixed structure. Newspaper also has multiple titles and multiple columns. In the proposed method, white gap area which separates titles, columns of text, line of text and words in lines have been identified to separate document into various segments. The proposed algorithm has been successfully implemented and applied over a large number of Indian newspapers and the results have been evaluated by number of blocks detected and taking their correct ordering information into account.*

Keywords— Document Layout Analysis, data extraction, document page segmentation, unstructured document.

I. INTRODUCTION

Document image segmentation is an important step in a document understanding system. The main aim of a page segmentation process is to separate a document image into its regions such as text, tables, images, drawings and headings. There are three different methods for page segmentation and layout analysis viz., a) Top-down, b) Bottom-up and c) Hybrid methods [11]. Top-down method start by detecting the highest level of structures such as columns and graphics, and proceed by successive splitting until they reach the bottom layer for small scale features like individual characters. In top-down methods, a priori knowledge about the page layout is necessary. Bottom-up methods start with the smallest elements such as pixels, merging them recursively in connected components or regions of characters and words, and then in larger structures such as columns. They are more flexible but errors are accumulated in every iteration. It makes use of methods like connected component analysis [12], run-length smoothing [13], region-growing methods [14], and neural networks [15]. Most of these methods are computationally expensive Many other methods are there that do not fit into top-down and bottom-up categories and therefore are called hybrid methods. Among these methods are texture-based [16] and Gabor filter [17]. Some work has done based on pyramid segmentation.

II. LITERATURE REVIEW

In the method [2] for segmentation and classification of digital documents using run length smearing approach. A linear adaptive classification scheme used to discriminates text regions from others. [3] proposed X-Y cut page segmentation approach based on top down approach. In this method entire document is considered as root and respective decomposed rectangular regions as leaf nodes. Horizontal and vertical projection profiles of foreground pixels are used for decomposition. [4] Proposed a method for segmentation of document using white space analysis method. In [4] segmentation technique is independent of any threshold values in proposed method. In these method white spaces runs greater than one fifth of the page are identified in both horizontal and vertical directions. Thinning algorithm is used for thinning of lines. In this way a mesh is formed by combining lines in both horizontal and vertical directions. [5, 6] proposed bottom-up approaches for page segmentation and block identification. The method [7] uses edge information to extract textual blocks

https://dx.doi.org/10.22161/ijaers.4.5.13 *ISSN: 2349-6495(P) | 2456-1908(O)*

from gray scale document images. The method detects only textual regions on heavy noise infected newspaper images and separate them from graphical regions. [8] Developed a new approach for page segmentation and classification based on white tiles. In this method, white tiles of each region have been collected together and their total area is estimated, and regions are classified as text or images. The method in [9] starts from pixel level information and finds k-nearest neighborhood pixels and start converging them. In this method, Document Image Segmented using dynamic thresholds and identification of thresholds are based on properties of distance and angle of each connected components with k-nearest neighbors. [10] proposed X-Y tree method which segments a document in multiple steps into a tree structure consisting nested rectangular blocks. In this method document is segmented into big blocks through horizontal or vertical cuts and then the same process proceeds in each one of the sub-blocks. The characteristic of the method is that its tree structure that suggests a logical order of the document. To reduce the computing cost modified methods were proposed.

From the above discussion, it is evident that there are many methods proposed in the literature to address the problem of document layout analysis. However, very few methods can be found in the literature for the analysis and segmentation of unstructured document images. Hence, in this paper, we propose a new method of segmenting unstructured document image. The proposed method, segment images occurring in a document and then we separate the headings in the document. After this we separate the columns and finally divide the text into paragraphs, lines and then words. Rest of the paper is organized as follows: Section 3 presents the proposed method. In section 4, we present the results of the proposed method and finally, conclusions are drawn in section 5.

III. PROPOSED METHOD

In this section, we present the details of the proposed method which address the problem of segmenting the unstructured document image. The method follows the top-down approach, by initially considering the complete document and trying to recursively extract text blocks from it in hierarchical way.

Our method can be categorized into the following categories

- Binarization of input image.
- Headline Segmentation.
- separating the columns
- separating the images and lines from the columns.
- Finally separating the words from the paragraphs

We discuss the above steps in detail in the following sections.

3.1 Binarization:

Scanned news paper document may be colored or grayscale image that needs to be converted to binary image to reduce the computational cost and helps to utilize simplified analysis methods. During thresholding logical and semantic content of the document need to be preserved for better understanding of the document image. Documents considered in our experiments are even colored paper, hence global thresholding algorithm Otsu[18] was used for binarization. An example of binarization is shown in figure 1 (a)


Micromax plans to enter top 5 global mobile cos list

Bigger Plans

MICROMAX RANKED 10th globally with 1.8 per cent share of the global handset market at the end of March 2015 and is second largest in India, says research firm Gartner.

Fig 1(a): Original Document

3.2 Headline Segmentation:

In news paper documents, text blocks in heading and text blocks in columns are characterized by thick of white pixels separated by block of black pixels as

(b) Binarized Image

shown in the fig 2. Height of the headings and columns of document can be found from the horizontal projection profile as shown in the fig 3. Using the height headings and columns of text from the document can be separated as shown Fig 4. Words in the heading are segmented based on the block of black pixels represented by white strip show in the Fig 5. Using vertical projection profiles width of the word can be found as show in Fig 6. Based on the project

https://dx.doi.org/10.22161/ijaers.4.5.13 *ISSN: 2349-6495(P) | 2456-1908(O)*

profile data, words are separated from the headings as shown in fig 7.

Fig 2 Heading and column text separated by white pixels

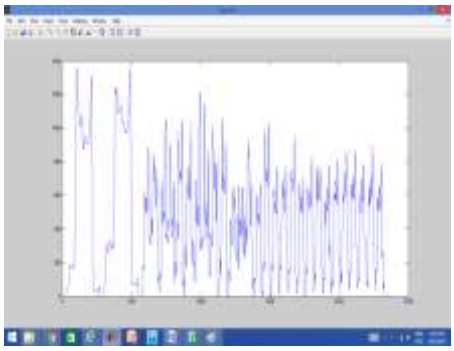
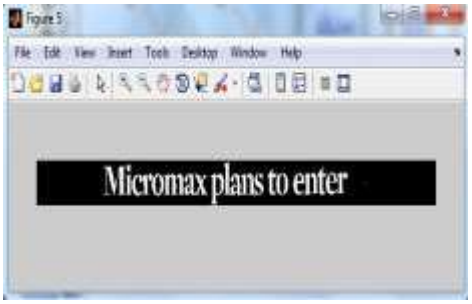
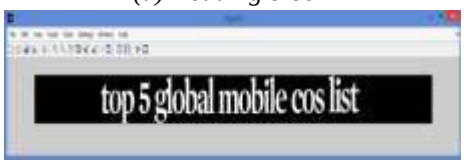


Fig. 3 Horizontal projection profile



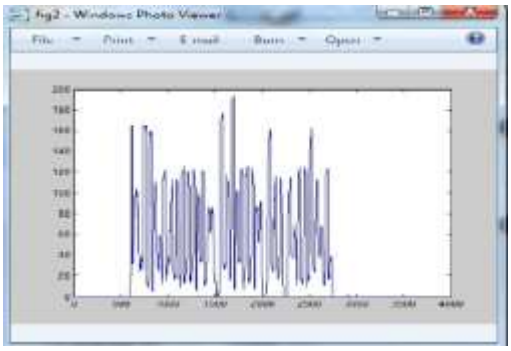
(a) Heading block



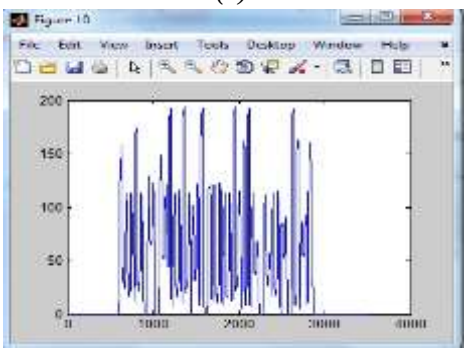
(b) Heading 2 block

Fig 4 (c) Column Block

Fig 5: words in the heading seperted by block of black pixels represented by white strip



(a)



(b)

Fig.6(a): Vertical projection profile of the heading1(b) Heading2

Fig.7: segmented words from heading1

3.3 Separating Columns from the Document: After the headers get separated from the images. The columns are separated, for this we use the width of the maximum occurring black run which separates the columns as shown

https://dx.doi.org/10.22161/ijaers.4.5.13 *ISSN: 2349-6495(P) | 2456-1908(O)*

in the Fig 8. It can be understood that the columns can occur only in those areas where the black spaces occur in large vertical blocks. We find the width of the black run using vertical projection profile as shown in the Fig 9.

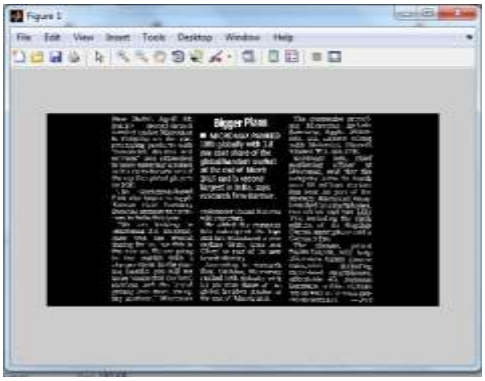


Fig 8: Columns seperated by black run

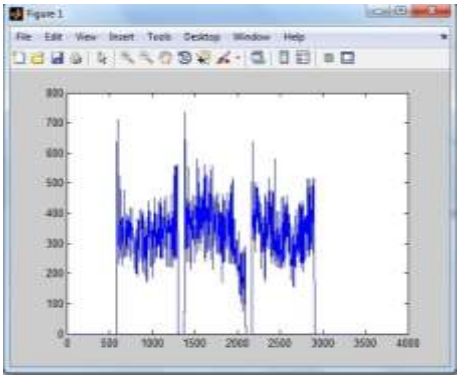


Fig. 9: Vertical projection profile of the columns

3.4 Separating images and paragraphs from the column:
In this step, we differentiate between text and images present in images. We retain only those runs of black pixels of the input image where the width of the black run is greater than or equal to twice the maximum occurring black run which is found empirically. We then use the connected component algorithm to find the height and width of these components. The components whose height is greater than a fixed threshold of 300 pixels can be said to be an image and hence removed from the document. We also consider the fact that if height is greater than a components width we call it an image and hence eliminate it from the document. We scan horizontally from each of the vertical lines to find continuous black pixel rows. We only consider those rows of black pixels where it can go horizontally till to end of the column using horizontal projection profile of the column as show in Fig 11; this is the threshold for segmenting lines from the columns as shown in the Fig 10.

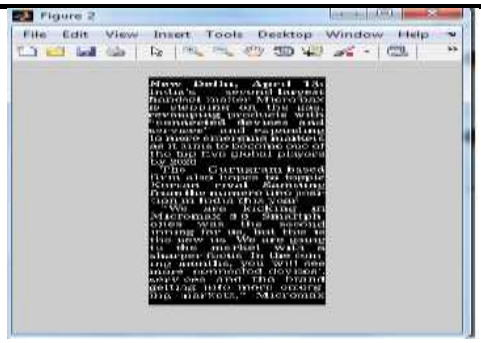


Fig 10: Text in the column

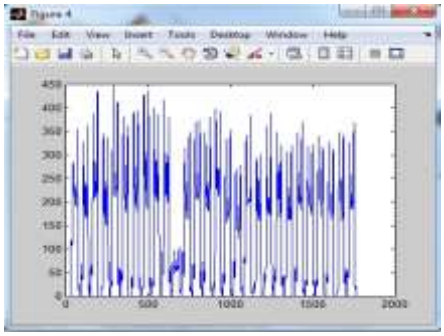


Fig 11: Horizontal projection profile

Separating words from the lines of the column: We use same method as mentioned in the first step. Using vertical projection profiles words from the lines of text are separated.

IV. RESULT AND DISCUSSION

To corroborate the efficiency and effectiveness of the proposed method, we have conducted extensive experimentation with large number of images. The images are obtained by scanning large number of English daily newspapers. Some of the sample outputs are shown below.

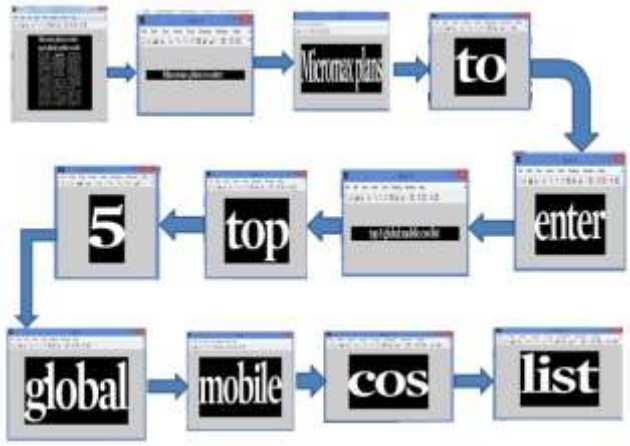


Fig.12: out of proposed method.

https://dx.doi.org/10.22161/ijaers.4.5.13 *ISSN: 2349-6495(P) | 2456-1908(O)*

The proposed method also handles large number of columns and multifont documents also. The Table 1 shows our testing results for 20 magazine pages.

Table.I:. Result for layout analysis

No of Pages	No of Blocks	Wrongly Segmented blocks	Wrongly merged block
25	200	20	10

The Fig 12 shows that the algorithm can accurately identify textual regions on english news paper document. These segmented blocks can be passed on as input to an OCR system.

V. CONCLUSION

In this paper, we presented an efficient algorithm for segmenting unstructured document images. For the purpose of evaluating the proposed method, we have conducted experiments on the English newspaper documents. We tested this algorithm on variety of documents from different newspapers with different page layouts. From the results of the proposed method, it is evident that the proposed method handles the complex and most challenging cases. The result also shows that, the proposed method is efficient and works in hierarchal form.

REFERENCES

[1] S. Khedekar, V. Ramanaprasad, S. Setlur, Text - Image Separation in Devanagari Documents. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003.

[2] K.Y. Wong, R.G. Casey and F.M. Wahl, Document analysis system. IBM Journal of Research and Development 1982; 26(6):647-656.

[3] B. Kruatrachue, N. Moongfangklang and K. Siriboon, Fast Document Segmentation Using Contour and X-Y Cut Technique. International Journal of Computer, Information science and Engineering 2007; 1(5).

[4] R. Garg, G. Harit and S. Chaudhury, A hierarchical analysis scheme for robust segmentation of Document Images using white-spaces. Proceedings of 1st National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics 2008. Document Image Segmentation Using Dynamic Thresholds and Identification 1875

[5] K. Lee, Y. Choy, and S. Cho, Geometric Structure Analysis of Document Images: A Knowledge- Based Approach. IEEE transactions on Pattern Analysis and Machine Intelligence 2000; 22(11).

[6] P. Mitchell, H. Yan, Newspaper document analysis featuring connected line segmentation. Proceedings of International Conference on on Document Analysis and Recognition, ICDAR'01, 2001.

[7] Q. Yuan, C.L. Tan, Text Extraction from Gray Scale Document Images Using Edge Information. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR'01, 2001: 302-306.

[8] A. Antona copoulos and R. T. Ritchings "Segmentation and Classification of Document Images", The Institution of Electrical Engineers 1995.

[9] L. O'Gorman, The Document Spectrum for Page Layout Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 1993; 15(11): 1162–1173.

[10] S. S. G. Nagy and S. Stoddard, Document analysis with expert system. Proceedings of Pattern Recognition in Practice II, June 1985.

[11] O. Okun, D. Doermann, and M. Pietikainen, Page segmentation and zone classification: The state of the art In UMD, 1999

[12] A.Jain and B. Yu, Document representation and its application to page decomposition. IEEE trans. On Pattern Analysis and Machine Intelligence, 20(3):294 308, March 1998.

[13] F. Wahl, K. Wong, and R. Casey, Block segmentation and text extraction in mixed text/image documents. CGIP, 20:375 390, 1982.

[14] A. Jain, Fundamentals of digital image processing. Prentice Hall, 1990

[15] C.Tan and Z. Zhang, Text block segmentation using pyramid structure. SPIE Document Recognition and Retrieval, San Jose, USA, 8:297306, January 24-25 2001.

[16] D. Chetverikov, J. Liang, J. Komuves, and R. Haralick, Zone classification using texture features. In Proc. of Intl. Conf. on Pattern Recognition, volume 3, pages 676 680, 1996.

[17] A. K. Jain and S. Bhattacharjee, Text segmentation using gabor filters for automatic document processing. Machine Vision and Applications, 5(3):169 184, 1992.

[18] M. Sezgin & B. Sankur (2004). "Survey over image thresholding techniques and quantitative performance evaluation". Journal of Electronic Imaging. 13 (1): 146–165.