Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 9; April-June, 2015 pp. 12-15

http://www.krishisanskriti.org/acsit.html

Document Layout Analysis for Hindi Newspapers

Shweta Singh[1], Tushar Patnaik[2] and Sneha Choudhary[3]

[1,2,3]CDAC, Noida, India
E-mail: [1]singh.shweta2721@gmail.com, [2]usharpatnaik@cdac.in,
[3]sneha.choudhary0106@gmail.com

Abstract—*Document layout analysis plays a vital role in automated document recognition system. Document layout retention is actually a preprocessing step to OCR. Accuracy of OCR decreases in case of complex layouts due to multiple columns, graphics, paragraphs, font-size, and text block properties. To increase the accuracy of OCR, we need to analyze layout of complex documents because automatic analysis of complex documents is still in nascent form. Document layout helps in identifying, categorizing and labelling the semantics of text/non-text blocks for meaningful information retrieval from document images. Our primary target document includes various Hindi newspaper images which are having complex layouts. The proposed approach is based on bottom-up methodology for analyzing the layout of complex document images. The proposed scheme firstly extract bounding box at word level using connected component analysis. After this, horizontal and vertical merging of these bounding boxes is performed simultaneously based on the properties of bounding box. Non-text blocks have been classified after horizontal merging based on some properties like area. Finally, reading order determination is performed using standard approach. The proposed approach have been successfully implemented and applied over a large number of Hindi newspaper pages. The accuracy has been evaluated by classification of text/non-text blocks, number of blocks detected and taking their correct ordering information into account.*

Keywords*: Layout retention, connected components analysis, preprocessing*

1. INTRODUCTION

The most important step of optical character recognition system (OCR) is document layout analysis.OCR performs conversion of scanned document images into editable text, but the efficiency of OCR decreases to very much extent in generating text data from complex document images. This is because the scanner fails to retain the order information of scanned document images. Hence, we can use document layout analysis effectively for automatic indexing and archiving of documents.

There are various techniques that have been used in past for analysis of layout. But most of the existing OCRs are still unable to retain the layout of complex documents images of Hindi newspapers. In our approach, we have used bottom-up methodology so that it can run faster and efficiently on complex document images.

This paper is organized in various sections. Section 2 describes the flow chart of proposed approach. This section explains the bottom-up methodology that has been used to generate block level data for document images. Section 3 describes the results of the approach. This approach has been successfully implemented on large number document image of Hindi newspapers

2. PROPOSED METHOD

In this section we present our approach of layout analysis which includes various modules. Fig. 1 shows an overview of the proposed approach.

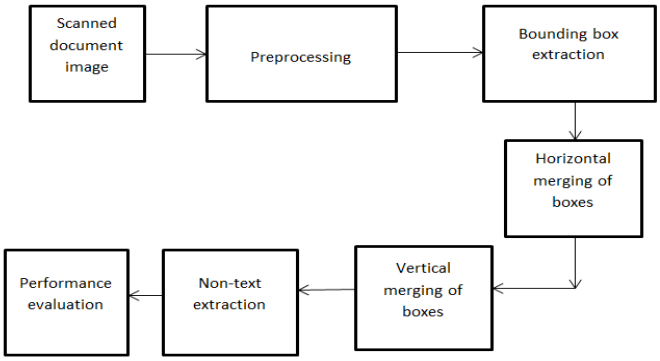

Fig. 1: Proposed approach

2.1 Preprocessing

At the very first stage, preprocessing of image is performed which includes image binarization, removal of noise and skew correction.For image binarization we have used thresholding technique, noise can be removed by various filters like mean/median filter and skew correction have been done by using hough transform.

2.2 Bounding box extraction

At the second stage of the proposed approach, extraction of bounding box at word level is performed. Hindi (Devanagari script) consists of special feature known as headline. This headline provides more connectivity in words as compared to other scripts.Therefore we have find out connected

components in the binarized image. Then, bounding box for each connected component is extracted. Fig 5 shows the result of this stage.

2.3 Horizontal merging

At third stage, merging of bounding box to get line level and block level box is performed. The word level bounding boxes are merged to give line level bounding box. These bounding boxes are merged according to properties of bounding box like height and width. If the height and horizontal inter-bounding box distance of two bounding boxes are compatible according to a given threshold value. Then, these boxes are merged in a single bounding box. In the similar way, we will get all the line level bounding boxes. Fig 6 shows the result of this stage.

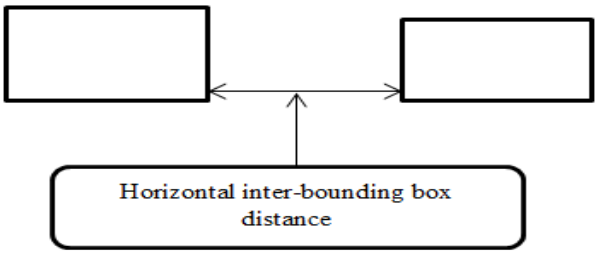

Fig. 2: Horizontal distance

2.4 Vertical merging

After horizontal merging, line level bounding box are merged to give block level box. For this, vertical inter- bounding box distance is calculated. If the vertical inter-bounding box distance between two boxes is compatible according to a threshold value, then these line level boxes are merged in a single block level box. In the similar manner, we will get all of the block level bounding boxes.Fig 8 shows the result of this stage.

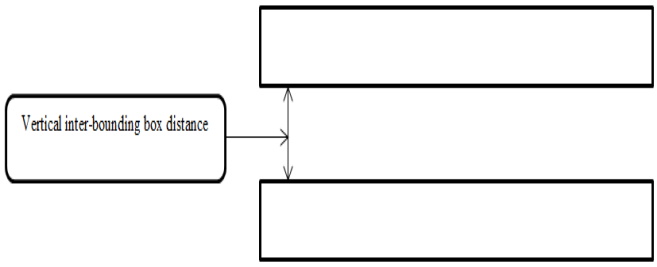

Fig. 3 Vertical distance

2.5 Non-text block extraction

At the fifth stage, text/non-text differentiation is performed. The two main factors that we have considered for differentiating between text/non-text are:

(a) Area of word level bounding box
(b) Aspect ratio of line level bounding box

Generally, the area of graphic region bounding box is much larger than the text region (word level) bounding box. Also, aspect ratio of graphic region is larger than that of line level bounding box. Aspect ratio of a bounding box can be calculated as:

Aspect ratio= Height of box/width of box

Therefore, after calculating area and aspect ratio of word level and line level bounding box respectively. We can extract the non-text (graphic) part from the image and store coordinate position of non-text block.

At the fifth stage, ordering of determined blocks has been performed. For block ordering standard approach have been followed for top to down and left to right reading sequence. Fig 7 shows the result of this stage.

2.6 block ordering

At this stage, ordering of determined blocks has been performed. For block ordering standard approach have been followed for top to down and left to right reading sequence. Fig 9 shows the result of this stage.


रीडर्स मेल

आरक्षण का आधार

आरक्षण आजादी के बाद से ही देश में बेहद जटिल मुद्दा रहा है। अभी हाल में उच्चतम न्यायालय ने जाट आरक्षण को खारिज कर दिया। पूर्ववर्ती संप्रग सरकार ने जाटों को आरक्षण देने की अधिसूचना जारी की थी। ओबीसी आरक्षण समिति ने इसको न्यायालय में चुनौती दी। फैसले में अदालत ने कहा कि जाति पिछड़ेपन का एक आधार हो सकती है, लेकिन एक मात्र आधार नहीं हो सकती। बेहतर हो कि सरकार पिछड़ेपन के आकलन के नए तरीके अपनाएं। मूलभूत प्रश्न है कि क्या देश के नीतिनियंता पिछड़ेपन के आकलन का नया पैमाना बनाने के लिए गंभीर हैं अथवा वोट बैंक की राजनीति ही आरक्षण का अंतिम सत्य है। लोकसभा चुनाव के ठीक पहले संप्रग सरकार ने केंद्रीय सेवाओं में जाटों को आरक्षण के दायरे में लाकर अनोखी चाल चली थी। आश्चर्य हुआ कि मौजूदा राजग सरकार ने भी इसके पक्ष में दलीलें पेश करने में पीछे नहीं रही। जाहिर है कि देश की कोई भी राजनीतिक पार्टी आरक्षण के खिलाफ कड़ा फैसला लेने का साहस नहीं दिखा सकती।

माना कि शोषित, वंचित एवं हाशिए पर रहने वाले वर्गों को विकास की मुख्यधारा में लाने का गंभीर प्रयास होना चाहिए। देश महाशक्ति बनने के मार्ग पर तभी अग्रसर हो सकता है जब समाज का सर्वांगीण विकास हो। महज कुछ लोगों के विकास से विकसित भारत का ख्वाब पूरा होने वाला नहीं है। परंतु क्या यह आरक्षण से ही संभव है? यकीनन आरक्षण से कुछ लोगों को फायदा मिला है, लेकिन इसकी देश को भारी कीमत चुकानी पड़ी है। राज्य एवं केंद्र सरकार की नौकरियों में ऐसे लोगों की बाढ़ आ गई है, जहां उनकी जगह नहीं बनती है। क्या कभी देश ने आरक्षण से होने वाले नुकसान पर विचार-विमर्श करने की जरूरत समझी? सच्चाई तो यह है कि सदियों से हाशिए पर खड़े लोगों का जीवनस्तर सुधारने के लिए नहीं बल्कि सत्ता में बने रहने के लिए आरक्षण को हथियार की तरह प्रयोग किया गया। हमारे संविधान निर्माता दूरदर्शी थे। काफी विचार करने के बाद समाज की घोर असमानता को कम करने के लिए आरक्षण को आवश्यक समझा गया। परंतु यह अनंत काल के लिए नहीं था। उम्मीद की गई थी कि समय के साथ आरक्षण स्वतः समाप्त हो जाएगा अथवा अन्य विकल्प का निर्माण होगा। परंतु क्या ऐसा हुआ? जाट आरक्षण के संदर्भ में उच्चतम न्यायालय का हालिया फैसला हमें रास्ता दिखाता

Fig. 4: Scanned document image

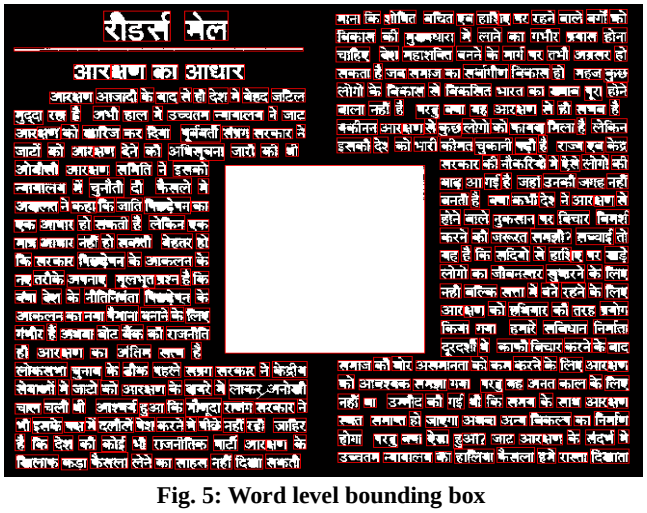

Fig. 5: Word level bounding box

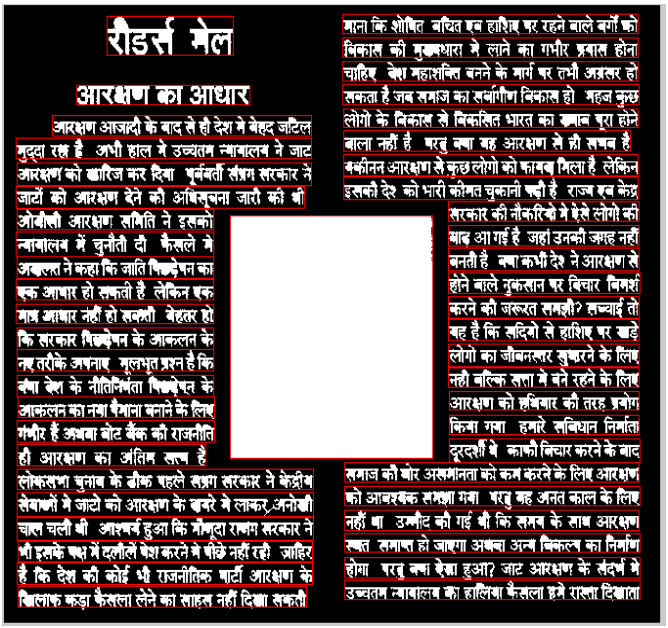

Fig. 6: line level bounding box

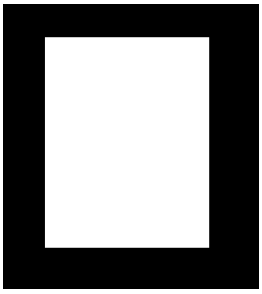

Fig. 7: Non-text extraction

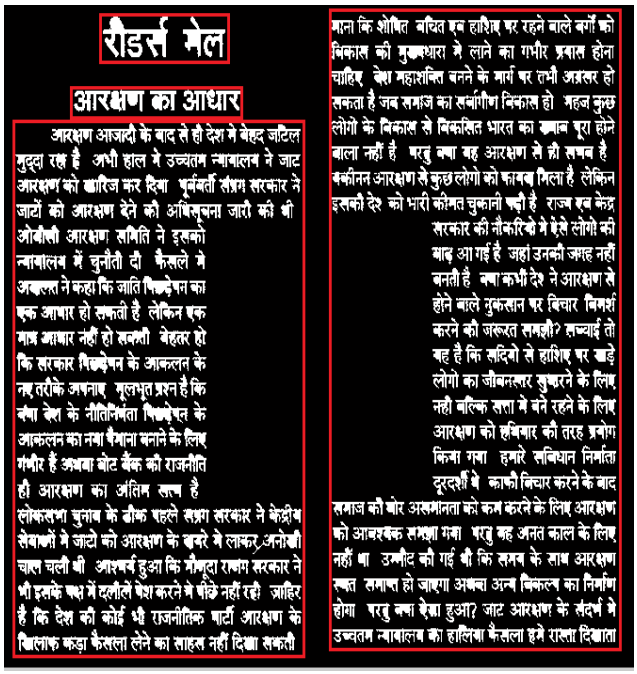

Fig. 8: Block level box

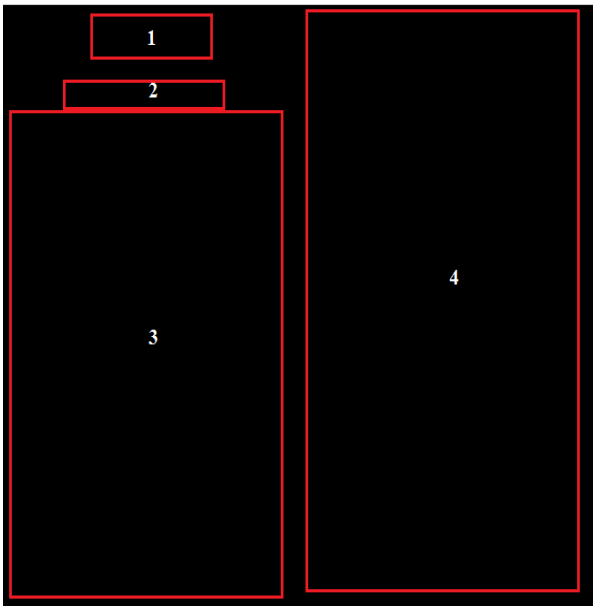

Fig. 9: Block Ordering

3. RESULTS AND CONCLUSION

This paper presents an effective approach to retain the layout of complex Hindi newspaper which contains multiple columns, multiple font size and graphics. This approach is able to identify the text/non-text regions in the document image of Hindi newspaper and also classify them. Finally, block ordering is done to determine order of extracted blocks.We have applied this approach on a large number of images of Hindi newspapers. The proposed methodology generates promising results.

The fig shows that the proposed algorithm can accurately identify textual/non-textual regions and can draw blocks on them. These ordered blocks can be passed on as an input to OCR system.

The approach works properly provided it follows the assumptions:

- Inter-row space between lines should not be greater than horizontal inter-block space.
- Inter-word space should not be greater than vertical inter-block space.

For the coming days, we will try to develop a generic approach for multilingual Indian newspaper documents so that it can effectively analyze the layout of any Indian script.

Table 1: Result of block detection

No of pages	Actual No of Blocks	Wrongly segmented blocks	Accuracy
40	415	16	96.14

Table 2: Block ordering results

No of pages	Actual No of Blocks	Wrongly Ordered blocks	Accuracy
40	415	65	84.33

REFERENCES

[1] A.RayChaudhuri,A.K Mandal, B.B Chaudhuri "page layout analyzer for multilingual Indian documents"Proceedings of the Language Engineering Conference (LEC'02) 0-7695-1885-0/12 © 2012 IEEE, 13-15 Dec. 2012.

[2] SupachaiTangwongsan,CholtichaBoondireke"A highly effective approach for document page layout extraction system"978-1-4799-2446-2/13©2013 IEEE

[3] F. Esposito,D. Malerba,G. Semararo"A knowledge-based approach to layout analysis"Proceedings of the fifth international conference on Document Analysis and RecognitionVolume:1 DOI: 10.1109/ICDAR.2005.599037 Publication Year: 2005, Page(s): 466 - 471 vol.1.

[4] KarimHadjar, Oliver Hitz and Rolf Ingold "Newspaper Page decomposition using a split and merge approach"0-7695-1263-1/01/© 2001 IEEE.

[5] J. Liang, J. Ha, and R.M. Haralick I .T. Phillips"Document Layout Structure Extraction Using Bounding Boxes of Different Entities"0-8186-7620-5/06 © 2006 IEEE.