Journal of Computational Linguistics
Volume 1, Issue 1, 2011, pp-05-10
Available online at: http://www.bioinfo.in/contents.php?id=46

Layout Segmentation of Scanned Newspaper Documents

A.Bandyopadhyay, A. Ganguly and U.Pal
CVPR Unit, Indian Statistical Institute 203 B T Road, Kolkata, India.

Abstract: ***Layout segmentation algorithms found in published papers often rely on some predetermined parameters such as general font sizes, distances between text lines, presence of images and document scan resolutions. Variations of these parameters in real document images greatly affect the performance of these algorithms. In this paper we present a simple and novel approach for document page segmentation which are complex in nature (having more than one picture or header). In this paper we have dealt with the segmentation of a scanned document into images, headers, columns and finally into paragraphs. We first separate the image and fonts of greater size and then follow it up with column separation. Finally we divide it into smaller paragraphs.***

1. Introduction

Document page segmentation is an important processing in a document image understanding system. The goal of a page segmentation process is to separate a document image into its regions such as text, tables, images, drawings and headings. This paper deals with the scanned newspaper images where more than one column can occur as well as more than one block of related texts with multiple pictures which are generally rectangular in shape.

In previous works mainly three different techniques for page segmentation and layout analysis are followed. They are: top-down, bottom-up and hybrid techniques [7].

The top-down techniques 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. For this type of procedures, 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 may suffer from accumulation of errors. It makes use of methods like connected component analysis [5], run-length smoothing [10], region-growing methods [4], and neural networks [9]. Most of these methods require high computation.

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 [1] and Gabor filter [6].

Some previous works have also been done based on pyramid segmentation. Recently, Zhixin Shi et al. proposed a method based on Dynamic Local Connectivity Map (DLCM) for block segmentation.

In this paper we 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.

Our paper has been organised as follows:

Section 2 deals with the related works, Section 3 deals with our approach, Section 4 deals with results, Section 5 deals with future work and Section 6 is our conclusion.

2. Related works

There are many works which have been published. We have gone through almost all of them for reference and after that we decided to create our own approach to this problem of layout segmentation of scanned newspaper documents with multiple topics and multiple columns.

Among the published methods, the X-Y tree method by Nagy and Stoddard [2] is an approach which segments a document in multiple steps into a tree structure consisting nested rectangular blocks. The method presents an intuitive procedure in which a 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 its tree structure that suggests a logical order of the document. The drawback of the method is its assuming that the text blocks in document images are in rectangular shapes which are well separated by rectangular streams of background pixels.

Run-length smearing [10] is a popular method that has been used in combination with many other ideas. It is basically a simple image processing done as following. Along each scan-line of a binary image horizontally and/or vertically, it turns the colour of small background runs into the colour of foreground. The process ends up with a blurring effect that glues all the close foreground pixels together. Run-length smearing efficiently increases the sizes of foreground connected components. It serves as a foundation for many other methods such as projection profile,

connected component grouping, neural network and many background based methods including white stream method [8].

In a recent work Zhixin and Govindaraju [12] proposed a method of Dynamic Local Connectivity Map where they study the image properties at different levels of gray image i.e., from 0-255. At every threshold example binarization at 255 will fill entire image with foreground colour i.e., white. There is only one connected component in this case and it represents the entire document page as one block. This is actually considered as root of the document image. Starting from here, the connected components from the binarized DLCM at lower levels are finer partitions of the document image. This binarization of connected component procedure is performed by using binarization threshold values from 255 downward, and hence a partition tree is build using these images respectively. More so their method concerns segmentation of column blocks and graphics. With this in view the authors propose the following methodology which has provided an excellent layout segmentation of a scanned newspaper document.

3. Proposed Method

Our method can be categorised into the following categories:-

- Smoothing the input image for the detection of mostly occurring white and black pixel runs vertically.
- Separating images from text.
- Separating fonts of greater size.
- Separating columns in between the smaller texts.
- Finally segmenting the document into paragraphs.

We discuss these steps in details in the following sections respectively.

3.1 Smoothing the input image

This step is the first step towards our goal of segmenting the layout of a scanned newspaper image. We first convert the grey image into binary image. We then smooth the binary image using the foreground colour i.e., black. We perform horizontal smoothing in which we check that if the distance of two consecutive black pixels is less than 25 we then make the pixels in between these pixels black. We choose 25 out here accordingly after testing our database of images with different thresholds.

After this we again smooth the image using a 5 X 1mask.

We basically follow this to find out two things. First, is to find the width of the occurrence of the most frequently occurring band of black pixels vertically. Second is to find the width of the occurrence of the most frequently occurring band of white pixels vertically. These two important data are the main foundation of our segmentation. A smoothed image is shown in Fig 1(b).

3.2 Separating images

In this step we differentiate between the text and images present in the image. Our algorithm was both effective and efficient in determining multiple images in the scanned documents we have worked with till date.

We again smooth the image properly such that the text blocks become long black rectangles. We create another image of the same size as the input image filled with just white pixels. 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 we already calculated. 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 400 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.

3.3 Separating headings

This step is similar to the previous step where 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 we already calculated.

After this we use the connected component analysis to label these components and find its height and width. These height and width of a connected component leads us to its bounding boxes which we use for font separation.

Thus after this step the greater font sizes are kept in a bounding box. We merge the bounding boxes of text lines which are related, so that the related texts fall under the same bounding boxes.

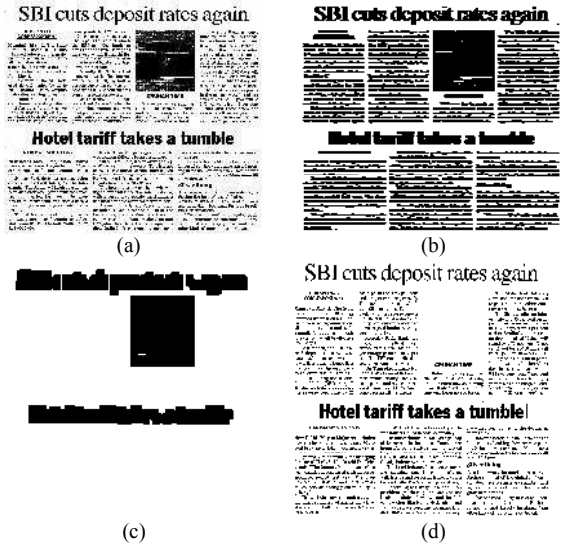


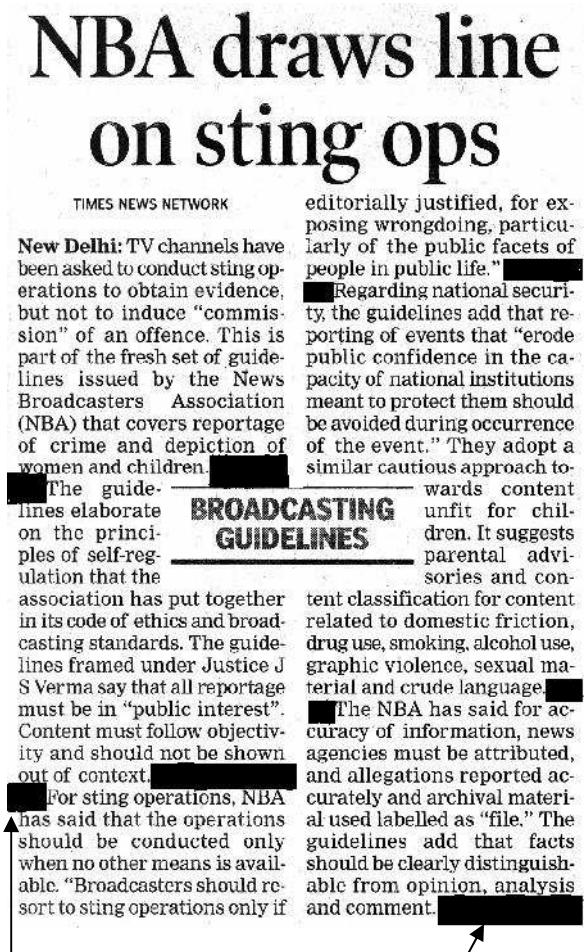
Fig 1. (a) shows a scanned document from a newspaper. (b) shows the image after smoothing. (c) shows the extracted componets of images and larger fonts. (d) shows the

document after the image is removed and the fonts of bigger sizes are kept inside bounding boxes.

3.3 Separating columns and objects occurring between columns

After the headers get bounded by the bounding box and the images get separated, we then separate the columns. We do this using the width of the maximum occurring white run which we have already calculated before in the first step.

We understand that the columns can occur only in those areas where the white spaces occur in large vertical blocks. We use a fixed threshold (which is 20 times the maximum occurring white run) to identify the columns. Since the columns occur consecutively for some pixels we choose only that pixel column which is the longest one in that column zone. We finally draw the column using the black pixel as a boundary in between two columns.



NBA draws line on sting ops

TIMES NEWS NETWORK

New Delhi: TV channels have been asked to conduct sting operations to obtain evidence, but not to induce "commission" of an offence. This is part of the fresh set of guidelines issued by the News Broadcasters Association (NBA) that covers reportage of crime and depiction of women and children.

The guidelines elaborate on the principles of self-regulation that the association has put together in its code of ethics and broadcasting standards. The guidelines framed under Justice J S Verma say that all reportage must be in "public interest". Content must follow objectivity and should not be shown out of context.

For sting operations, NBA has said that the operations should be conducted only when no other means is available. "Broadcasters should resort to sting operations only if editorially justified, for exposing wrongdoing, particularly of the public facets of people in public life."

Regarding national security, the guidelines add that reporting of events that "erode public confidence in the capacity of national institutions meant to protect them should be avoided during occurrence of the event." They adopt a similar cautious approach towards content unfit for children. It suggests parental advisories and content classification for content related to domestic friction, drug use, smoking, alcohol use, graphic violence, sexual material and crude language.

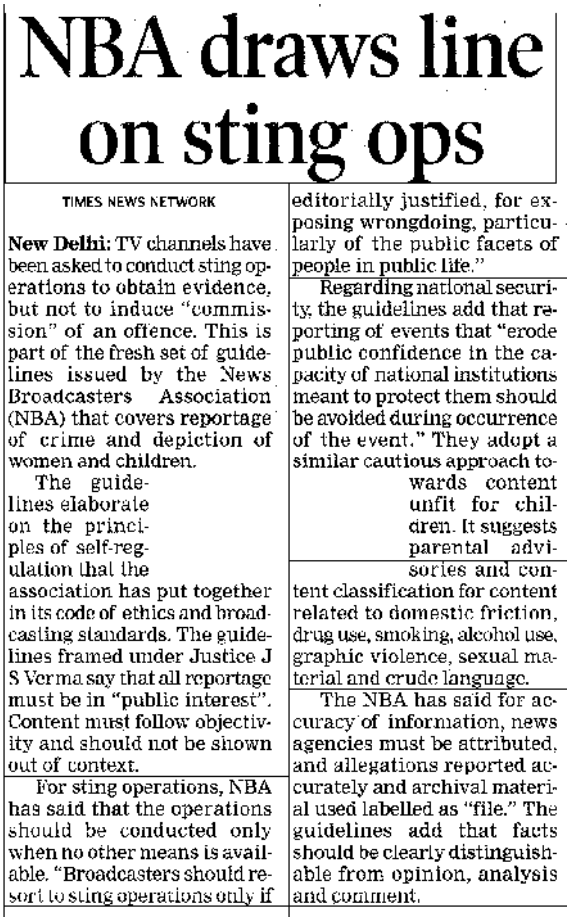
The NBA has said for accuracy of information, news agencies must be attributed, and allegations reported accurately and archival material used labelled as "file." The guidelines add that facts should be clearly distinguishable from opinion, analysis and comment.

BROADCASTING GUIDELINES

(a)Shows the starting of a paragraph (b)Shows the ending of a paragraph

We usually follow a certain rule while drawing the line which is, that the line we are going to draw should be the longest one in its band and it should be drawn at the last occurring index of its length. This will help in making the column lines nearer to the text blocks and thus help us later while separating paragraphs.

Many a times it can be found out that either texts or graphics occur suddenly in between two columns. This creates a problem of segmenting the columns properly. We have also taken care of this particular problem in our paper. An illustration of such an image is shown below in Fig 2.



NBA draws line on sting ops

TIMES NEWS NETWORK

New Delhi: TV channels have been asked to conduct sting operations to obtain evidence, but not to induce "commission" of an offence. This is part of the fresh set of guidelines issued by the News Broadcasters Association (NBA) that covers reportage of crime and depiction of women and children.

The guidelines elaborate on the principles of self-regulation that the association has put together in its code of ethics and broadcasting standards. The guidelines framed under Justice J S Verma say that all reportage must be in "public interest". Content must follow objectivity and should not be shown out of context.

For sting operations, NBA has said that the operations should be conducted only when no other means is available. "Broadcasters should resort to sting operations only if editorially justified, for exposing wrongdoing, particularly of the public facets of people in public life."

Regarding national security, the guidelines add that reporting of events that "erode public confidence in the capacity of national institutions meant to protect them should be avoided during occurrence of the event." They adopt a similar cautious approach towards content unfit for children. It suggests parental advisories and content classification for content related to domestic friction, drug use, smoking, alcohol use, graphic violence, sexual material and crude language.

The NBA has said for accuracy of information, news agencies must be attributed, and allegations reported accurately and archival material used labelled as "file." The guidelines add that facts should be clearly distinguishable from opinion, analysis and comment.

Fig 2. (a) Shows a scanned image from a newspaper where there is an intrusion in between two columns. In this figure we have highlighted the beginning and ending of a paragraph by black horizontal rectangles. These blocks will help us in separating paragraphs. (b) Shows the input image with the intrusion in between the column deleted and columns properly separated.

3.4 Segmenting the rest of text into paragraphs

We keep the addresses of the vertical lines which we draw to segment columns. Using these we try to segment the smaller blocks of texts into paragraphs.

We scan horizontally from each of the vertical lines to find continuous white pixel rows. We only consider those rows of white pixels where it can go horizontally till 40 pixels from our present vertical line; this is our threshold for segmenting paragraphs for the database of images we worked on.

We understand that the rows of white pixels even occur in between two consecutive lines of text. Hence, we filter them out by keeping only those groups of white rows whose combined width is greater than that of the width of the maximum occurring white run of pixels vertically which we already found out before. Since, the maximum occurring run of white pixels in a document is the width of white runs of pixels in between two text lines. We have marked these horizontal white rows in Figure 2(a) which will lead us for separating paragraphs. Therefore, any width of white pixel rows greater than this width can be easily said to be a starting of a paragraph, as we note that a new paragraph never starts from extreme left, instead, we leave a certain gap towards the left which is a general practise. This is how we separate individual paragraphs from texts. The image shown in Figure 3 is after the segmentation done on it by our proposed method on the input image shown in Figure1 (a).


SBI cuts deposit rates again

CRUNCH TIME

Hotel tariff takes a tumble

Silver lining

Fig. 3 shows a scanned document (whose input image is already shown in Figure 1(a)) after our segmentation.

4. Results and Discussions

For the research especially emphasized on finding feasibility of our method, we have chosen images from several different document image sources. These include document pages scanned from daily newspapers in two different languages English and Bengali scanned with 300 dpi resolutions. Figure 4 shows an example result of a scanned document in Bengali.

Subjective evaluations show that our method is a robust approach for segmentation of most printed documented images. Our research work is still continuing. The focus is on an efficient implementation of the algorithm on all possible type of document images and on all possible font sizes.

As of now our research is on dealing with the false connections between different document objects due to binarization noise. The possible solution might be using connected component to separate the characters.

We understand that the image can be of various shapes and can occur at any position of a document. We therefore have prepared our dataset in such a way that we face all these problems in our working.

We have classified our data mainly on two different respects one in English and one in Bengali. We have worked with 70 images and have got encouraging results using our method.

We have got some cases where we got over segmentation and under segmentation of paragraphs by our proposed method. In our case over segmentation of a paragraph is when a single paragraph gets separated in two or more text blocks. Similarly, under segmentation is when two paragraphs get treated as a single block of text.

The detailed results achieved by our method are shown below in Table1.

Table 1- Shows the results we achieved using our proposed method.

Languages tested	English	Bengali
Number of images tested.	30	40
Accuracy of column detection (in %)	100	97.6
Number of columns actually present	78	84
Accuracy of paragraph detection (in %)	99.35	97.77
Number of paragraphs actually present	312	182
Percentage of over segmentation in paragraphs	1.60	0.55
Percentage of under segmentation in paragraphs	0.96	1.66

The method is implemented using Microsoft Visual C++ and it is being run on a PC with Intel Dual-Core Mobile Processor 2.00 GHz. The processing time for each document varies with the size of the document. For images whose size is greater than 2000×2000 is about 5 seconds. The actual processing time depends on the data structure and platform used by the technique.

নয়া সরকারের কাছে কর কাঠামো সরল করার দাবি গাড়ি শিল্পের

Fig. 4 shows a scanned document in Bengali.

Some erroneous results were found while working with images in Bengali where two paragraphs were getting placed inside the same block. This was due to a problem in the threshold of those images.

In our paper we have dealt with images but graphics was not dealt with. In extreme cases images cannot be segmented if it does not match our criteria for its segmentation but such cases are rare. In those cases the images gets separated in a block which can later be segmented out.

Many a times it is found out that an image or a line of text occur in between two columns. In those cases our algorithm may suffer from the problem of over segmentation. This means that a paragraph may even get separated into another smaller block which actually is not there. An example of such an image is shown in Figure 5(b) and Figure 2(b).

Presently we faced a problem where there is a single text of different font which starts the document. We have eliminated those from the document as shown in Figure 5(a) where 'C' is of different sized and single and hence we eliminate it.

5. Future work

Our research started with a main motive of finally segmenting all texts blocks into paragraphs and its headings. Since our algorithm is not dependent on any particular type of language we plan to test it with all the Indian languages present in our country. We then can develop this as the first step towards the development of any OCR of any language which is present.

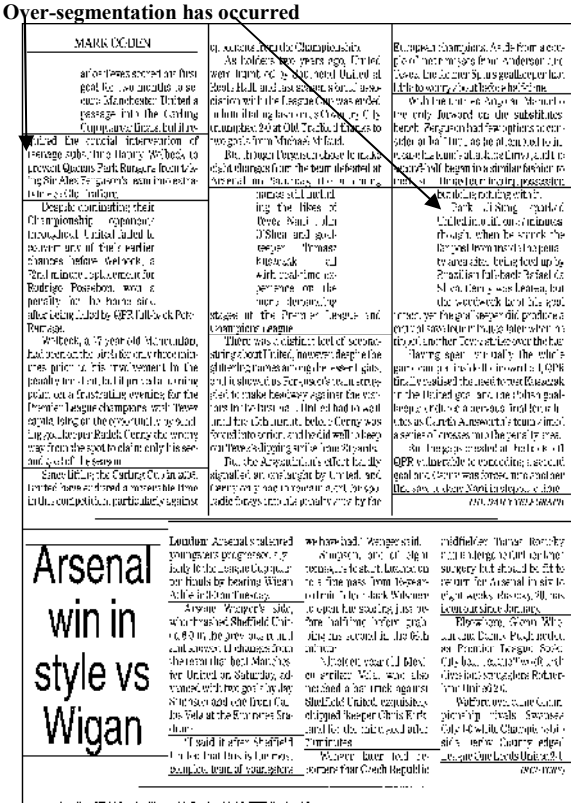
After this we plan to label each heading and its related component. This is a necessary step because doing this will make the algorithm more pronounced in using it for the first step towards the development of an OCR.

MARK OGDEN

Arsenal win in style vs Wigan

Fig. 5(a)

Over-segmentation has occurred



MARK OGDEN

Arsenal win in style vs Wigan

Fig. 5(b)

Fig. 5(a) is the input image.

Fig. 5(b) is the image after segmentation. The places where the images are removed an over segmentation has occurred.

6. Conclusion

In this paper we have described the major components of a complete text layout. We have approached the document analysis and document understanding in a new way. For the layout segmentation of a document, this paper presents a method capable of robust behaviour even for multi-columned documents with photographs and with multiple headers. This approach gives the items on the page a hierarchy which models the relationships between paragraphs, columns, headers, images and the page. For document understanding, this paper reports an attempt to build a method of understanding document layouts without the assistance of the character recognition results, i.e., the meaning of the contents. This paper aims at developing the segmentationwhich can be used on various languages.

Experimental results on a variety of documents have shown that the proposed methods are applicable to most of the document types commonly encountered in daily use, although there is still room for further refinement in the transformation rules for document understanding and resolution of ambiguities in noisy documents.

References

[1] 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.

[2] S. S. G. Nagy and S. Stoddard, "Document analysis with expert system". *Procedings of Pattern Recognition in Practice II*, June 1985.

[3] M. Hose and Y. Hoshino, "Segmentation method of document images by two-dimensional fourier transformation". *System and Computers in Japan*.

[4] A. Jain, "*Fundamentals of digital image processing"*. Prentice Hall, 1990.

[5] 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.

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

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

[8] T. Pavlidis and J. Zhou, "Page segmentation by white Streams". *Proc. 1st Int. Conf. Document Analysis and Recognition (ICDAR),Int. Assoc. Pattern Recognition*, pages 945–953, 1991.

[9] C. Tan and Z. Zhang, "Text block segmentation using pyramid structure". *SPIE Document Recognition and Retrieval, San Jose, USA*, 8:297–306, January 24-25 2001.

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

[11] D. Wang and S. Srihari, "Classification of newspaper image blocks using texture analysis". *CVGIP*, 47:327–352, 1989.

[12] Z. Shi and V. Govindaraju, "Multi-scale Techniques for Document Page Segmentation" pp.1020-1024, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005

[13] G.Harit, R.Garg and S.Chaudhury, "Syntactic and semantic labelling of hierarchical organized document image components of Indian Scripts". Proc. ICAPR, pp. 314-317, 2009.