# Page Layout Analyser for Multilingual Indian Documents

A. Ray Chaudhuri[♣] A. K. Mandal[♠] B. B. Chaudhuri[♣] *Fellow IEEE*
[♣]*Computer Vision and Pattern Recognition Unit*
[♠] *Electronics and Communication Sciences Unit*
*Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India*
*anirban@isical.ac.in, bbc@isical.ac.in*

## Abstract

*An advanced Optical Character Recognition (OCR) system is equipped with the module of the page layout analyser. It separates textual zones from non-textual zones. It identifies textual blocks from multicolumn documents and groups them into homogenous regions in terms of geometric shape and spatial distribution. All existing OCR modules developed for various Indian scripts can handle text only single-column documents. In this paper, a page layout analyser that uses typical common features present in most of the Indian scripts is introduced. A simple compatibility criterion that allows various degrees of homogeneity is defined. The page-analyser is robust in the sense that it can distinguish text regions from non-textual entities such as images, rulers, and noisy signals due to smudges and poor quality of the paper. Test results are shown in two most popular Indian Scripts, Devnagari (Hindi) and Bangla.*

## 1. Introduction

Transformations of paper document to its electronic version and subsequent document image understanding have become an important and challenging application domain. As a key module in document image understanding, *Optical Character Recognition* (OCR) has received the most attention for several decades. Nagy [1] is referred for an up-to-date review. A basic OCR system recognizes textual parts from a single column paper based document automatically and converts them into electronic format for further processing. *Page layout analysis* which is part of an advanced OCR system, segments or partitions the single or multicolumn document into several blocks including text blocks, which can be fed to the basic OCR module. As a result, the performance of the document understanding system as a whole greatly depends on the page layout analyser that precedes basic OCR. Several OCR software for complex documents in European Languages particularly in English are available that can handle multicolumn, images, tables, graphs and also able to preserve the document layout. On the other hand, the work on document analysis and recognition in Indian languages are still in a nascent form. A few basic OCR modules are available. Most of these modules from North Indian languages such as Hindi, Bangla, Punjabi, Assamese, are based on Brahmi scripts. These Brahmi based scripts have some common characteristics. Most of the characters have a horizontal line at the upper part called *headline* and primarily the characters of words in these scripts are connected by these headlines.

In this paper, we report the progress of the ongoing project on page layout analyser for Brahmi script based documents where headlines dominantly exist and present results for two most popular Indian scripts viz., Bangla and Devnagari (Hindi). The main textual blocks in typical Bangla or Hindi documents (of books, newspapers, magazines etc.,) are in single, two or three columns style. In addition, the texts are also often present in the title, headings, authors name and their affiliation, footnotes, page numbers, and figure-captions. Few occasions, some table-like structures also exist in the documents but such structures are not considered here. The texts are printed in various fonts, sizes, styles and spacing whereas paragraphs are formatted with different indentions and justification rules.

At present, the analyser is capable of localizing and extracting text regions from binary images of documents, as rectangular blocks (intersecting or disjoint) containing moderate sized connected components satisfying *homogeneity or compatibility* in terms of headline and other (textual) features particularly available in these Indian scripts (details are presented in Section 3.1). A simple compatibility criterion, which allows various degrees of homogeneity in terms of selected features, is applied. The generation of textual blocks is established by taking the optimal *bounding boxes* (BBs) of all connected components in the converted two-tone image from the

Proceedings of the Language Engineering Conference (LEC'02)
0-7695-1885-0/02 $17.00 © 2002 IEEE


**Figure 1. Upper, Middle and Lower Zoning (UMD) of Bangla and Devnagari text-lines.**

original gray scale image. Among these BBs, a few having headlines are taken as *seeds*. These seeds are grown along vertically overlapped and compatible BBs. Refinement and corrections of initially generated text-blocks are also performed. The page layout analyser is robust in the sense that it can distinguish text regions from non-textual entities such as images, rulers, and noisy signals due to smudges and poor quality of the paper. The major part of the computation is on the feature domain rather directly on the image and the method is considerably fast.

We begin the following section by elaborating the problem of page layout from geometric and spatial distribution standpoint. Existing popular techniques on page layout analysis for European documents, particularly on text-region identification are briefly mentioned. In Section 3, the results of the analysis required for deriving the feature set for discriminating textual zones in both Bangla and Devnagari scripts, are summarized. The basic compatibility criterion used for finding homogeneity among the blocks is also presented. In Section 4 the implementation details are presented. Experimental results are provided in Section 5. Section 6 summarizes the work and indicates the directions of the future work.

## 2. Page Layout Analysis: The Problem

The logical comprehension of arbitrary documents is a complex task involving high-level mental processes, simulation of which are currently out of the capability of an automatic system. A good target is to perform automatically the *Geometric Layouting* or *Zoning* [2]. It aims at producing a description of the geometric structure of the document. It starts with some pre-processing including image acquisition, skew-correction and image noise reduction. Then the *page decomposition* step labels the document image into homogeneous blocks of maximum size (this step is called *page segmentation*), and to classify them into a set of predefined data types (this step is called *block classification*). Page segmentation takes into consideration only the geometric layout of the page e.g., the spacing among different regions, while block classification employs specific knowledge about the data types to be discriminated, e.g., features can be derived to distinguish among text, images, or drawings. Among the pre-processing modules, an important one is that of skew correction. It estimates the document orientation angle with respect to the horizontal and rotate the image to the desired orientation of zero degree. In our project, an efficient graphics user interface based technique is developed which can handle any document with arbitrary skew. The detail of the technique is described elsewhere. Here it may be assumed that the amount of skew is nominal in all documents.

### 2.1. Existing Page Layout Techniques for Text-Region Recognition

Any page layout analysis technique depends on few factors such as the language (script), style, subject domain etc. There are several types of page layout analyser available for European scripts particularly for English. However, to the best of our knowledge no published work on page layout analysis for documents on any Indian languages are available. Note that character level connectivity in words exists in north Indian Brahmi-based scripts. On the other hand, in south Indian scripts (based on Grantha script), primarily words are formed by disconnected characters due to absence of headlines. Thus, from page layout perspective, scripts in south Indian languages are globally much closer to English than north Indian scripts and existing page layout analysers developed for English documents may be applied directly or with little modifications to south Indian scripts.

A few popular methods used in English and similar European languages for text-region recognition are mentioned below.

**Smearing based techniques:** One of the first approaches to text/non-text discrimination was presented by Johnston [3]. It was assumed that the image is clean with known character height-width in pixels, horizontally aligned texts etc. The algorithm works in two steps: (1) Clean-up objects larger than the character size by applying some morphological operators whose parameters depend on the character size. (2) Clean-up objects smaller than the character size, using similar morphological operations.

**Connected components analysis:** O'Gorman [4] described a method that aimed at locating text lines and text blocks in images containing text, tables or equations. It was based on the computation of the document spectrum, called *docstrum*. A pre-processing phase was performed to remove noise. The connected components were then clustered according to their areas. Connected component based several other methods are available in the literature such as Jain and Yu [5].

**Projection profile methods**: A simple method to segment textual documents into lines is based on the

analysis of the regularity of peaks and valleys in the projection profile of the image performed along the text orientation. Work by Srihari and Govindaraju [6] and Pavlidis and Zhou [7] may be consulted for details.

**Texture based or local analysis:** Chen *et al.* [8] described a segmentation algorithm for the detection of words in textual documents. A word block was defined as the rectangular regions, which stored a word. Word blocks were detected by means of a pixel classifier, which gives an estimate of the probability of the pixel to be in a word block or not.

**Analysis of the background structure:** Baird [9] described a segmentation technique in which the structure of the document background was analysed. In a pre-processing phase, the components that appear too small or too large in the text were filtered and the document was de-skewed. All the maximal rectangles, *i.e.,* white rectangles, which could not be further expanded, covering the background, were enumerated. A partial order was specified according to the area and the aspect ratio of the rectangles. The regions not covered by the background constitute the blocks. The advantage of this method is that it does not rely on rules dependent on the characters set but only requires a rough estimate of the range of text size.

## 3. Scripts Characteristics: Derived Features and the Compatibility Criterion

All major north Indian scripts including Bangla and Devnagari are mixtures of syllabic and alphabetic scripts and derived from Brahmi script through various transformations and most of the characters have headlines. Compared to English, the font and style variations in these scripts are few. Popular fonts for mass printing media are still Monotype and Linotype. A text line in these scripts may be partitioned into three horizontal zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic (and compound) characters below headline and the lower zone is the portion where the modifiers can reside. This typical zoning, called *UMD* (Upper Middle Down) is shown in Figure 1. If several consecutive characters of a word have headlines, they make a continuous straight line, which, if properly detected, can be used as a discriminating feature for text block extraction. It can be easily established that textual blocks in either Bangla or Devnagari are 'well covered' by headlines or in other words, almost all words contain headlines.

**Likelihood of headline in a word:** In Bangla, 41 characters can appear in the first position of a word. Out of these, 30 have headlines. Hence probability of getting a character with headline in the first position of a word, $p_1 = 30/41$. Characters, which can contribute to the headline in the other positions of a word, are mostly consonants. Since 28 out of 39 Bangla consonants have headlines, the probability of getting a consonant with headline for other positions in a word, $p_2 = 28/39$. Then probability of all four characters without headline in a word is $(1- p_1)(1-p_2)^3 = 0.00601$ (assuming that all characters are equally likely and independently occurring). Hence, probability that a word will have at least one character with headline is 0.99399. Analysing in the same way we get for Devnagari, the corresponding probability of 0.99702. The practical situation is better than these estimates since characters are not equally likely in a word and most frequently used characters have headlines [10].

**Likelihood of vertical line in a word:** Another important hypothesis is that most of the words have vertical lines that spread along M, the middle zone of UMD. Justifying as above, the probability of existence of such vertical lines in a word are 0.9796 and 0.9845 respectively in Bangla and Hindi. The position of the vertical line with respect to its BB provides the length-wise break-up in U, M and D of UMD.

### 3.1. Primary and Secondary Discriminating Features

The features extracted from BBs viz., corners, headline (position and its thickness), the vertical line in the middle zone of UMD, lengthwise distribution of UMD and average number of object pixels per BB, i.e., density of the component are used as *primary discriminating features* for text block identification. At the later stage of computation for refinement and corrections of initially generated text-blocks, a few *secondary features* or *block level features* are derived from the primary feature-set. The average pixel density of each component, average vertical span of M within a block, average horizontal gap among closest pairs of BBs, inter-headline distances between adjacent vertical BBs, inter-block distances (both horizontal and vertical) and the ratio of number of components between adjacent blocks are effectively used.

### 3.2. Compatibility Criterion

For any two scalar quantity $L_1$ and $L_2$ the compatibility is defined as

$$Comp(L_1, L_2) = \frac{1}{2}\left(\left|1 - \frac{L_1}{L_2}\right| + \left|1 - \frac{L_2}{L_1}\right|\right)$$

If $Comp(L_1, L_2) < \varepsilon$, a given small fraction taken as threshold then $L_1$ and $L_2$ are said to be mutually $\varepsilon$-compatible.

By varying the value of $\varepsilon$, the degree of compatibility is established and this measure is used in all compatibility checking.

राज्य पुलिस की कार्यशैली पर अंगुली उठी

नगर प्रतिनिधि

कोलकाता, ६ जुलाई। पुलिस कर्मियों की संख्या अधिक कहां बढ़ायी जानी चाहिये? इसका सीधा उत्तर है जहां अपराध अधिक हो लेकिन अगर २००१ के आंकड़े को देखें तो बीरभूम के तत्कालीन पुलिस अधीक्षक ने अपराध अधिक होने वाले स्थानों पर कम पुलिस तैनात किया था जबकि जहां अपराध कम होते हैं वहां अधिक पुलिस कर्मियों को तैनात किया गया। भारत के नियंत्रक व महा लेखा परीक्षक ने अपनी रपट में इस बात को कहा है। इसके लिउए उदाहरणस्वरूप बीरभूम को दिखाया गया है। यहां के आंकड़े बताते हैं कि अपराध के मामले में सिउड़ी इस जिले में प्रथम स्थान पर है लेकिन सिउड़ी थाने में केवल ४१ पुलिस कर्मी रखे गये हैं। इसके विपरीत रामपुरहाट जहां तुलनात्मक रूप से अपराध कम हुआ है वहां ५० पुलिस कर्मियों को रखा गया। रामपुर हाट में कुल अपराध ७१ हुए वहीं सिउड़ी में कुल अपराधों की संख्या १४९ है। इसके साथ ही कैग ने एक और महत्वपूर्ण बात को ओर अंगुली उठाई है। यह महत्वपूर्ण बात यह है कि महिलाओं के खिलाफ अत्याचार की घटनाओं के बढ़ने के बावजूद महिला पुलिस की संख्या में कोई खास वृद्धि नहीं हुई है।

बीरभूम को ही उदाहरण के रूप में दिखाते हुए कहा गया है कि १९९६ से १९९९ तक केवल ९ महिला पुलिस इस जिले में थी। २००० में इनकी संख्या बढ़ाकर २१ की गयी। कहने का अर्थ यह है कि राज्य में जिस समय ९७५ महिला पुलिस कर्मी थीं उस समय बीरभूम में इसका केवल १ प्रतिशत यानी ९ महिला पुलिस कर्मी थीं, जबकि महिलाओं के खिलाफ अपराध १४ से १८ प्रतिशत बढ़े। कैग की रपट में राज्य पुलिस के कर्मचारियों की संख्या कम होने का भी उल्लेख किया गया है। इसमें बताया गया है कि प्रत्येक पुलिस कर्मचारी को ९.३१ किमी. तथा ६०९२ की आबादी कवर करनी पड़ती है। अपराध बढ़ने का एक कारण यह भी है। नेशनल क्राइम क्रिमिनल इंफारमेशन सिस्टम को राज्य सरकार ने सही तरीके से लागू नहीं किया है। दो करोड़ रुपये क्राइम क्रिमिनल इंफारमेशन सिस्टम को १९९४-९५ में शुरू किया गया था। इसका उद्देश्य अपराध की बढ़ती प्रवृत्ति और संगठित अपराध के संबंध में सभी राज्यों को बताने का था लेकिन आन लाइन कनेक्टिविटी नहीं मिलने के कारण राज्य में लगभग सात वर्षों से यह प्रणाली शुरू ही नहीं हो पायी। कैग ने इसके लिए राज्य स्तर पर मानिटिरिंग, सहयोग का अभाव, अपूर्ण प्रशिक्षण को जिम्मेदार ठहराया है। क्राइम क्रिमिनल इंफारमेशन सिस्टम को लाने के लिए केन्द्रीय गृह मंत्रालय ने राज्य सरकार को अलग से गैर योजना कोष दिया था। इसके अलावा नेशनल क्राइम रिकार्ड ब्यूरो ने नवम्बर १९९४ से अगस्त २००० के मध्य ४९ कम्प्यूटर मुहैया करवाये थे। इनमें कई कम्प्यूटर तो सिर्फ रखे हुए ही हैं। इस प्रणाली के लिए इनका उपयोग नहीं किया जाता। राज्य सरकार को इस प्रणाली के लिए गैर योजना कोष में १५.४८ लाख रुपये दिये गये थे। इसमें १४.७६ लाख रुपये ही खर्च किये गये।

**Figure 2. A multicolumn document taken from a Hindi newspaper *Sanmarga*.**

## 4. Implementation

### 4.1. Image acquisition and binarization

To scan the image a flatbed scanner is used and digitised images are stored in a 8 bit gray scale TIFF format at 300 dpi. It is assumed that images are bimodal and the binarization is done by thresholding. Any Pixel with a value less than a given threshold $T$, set to one (object); otherwise, set to zero (background). The default threshold is 128.

### 4.2. Component labelling and bounding box computation: extraction of primary features

The component labelling is performed by a recursive neighbourhood search and tagging. The top-left and the bottom-right corners of the BB and the component density are directly computed from the rectangular span of the connected component.

**Headline positioning:** In the upper portion (≈ 40% of height) of the BB, consider the row with maximum number of object pixels. If the number is greater than 60% of the height of the BB and the maximum number of consecutive object pixels in that row is greater than a threshold then that row defines the possible existence of a headline and the row number is the headline position. If no such row exists, then the BB is considered to be without any headline.

**Headline zoning:** To find the headline zone, a growing process along the upper and lower side of the headline position is performed. As long as the total number of object pixels in the immediate adjacent row of the headline zone is compatible with that of the headline position, the row is appended to the headline zone. Finally, the topmost compatible row-position indicates the headline-start whereas the lowermost compatible one, the headline-end. If the thickness of the headline zone is more than $1/8^{th}$ of the height of the BB or headline-end position resides below the designated area of upper portion of the BB then that BB is also tagged as without any headline.

**Upper, middle and lower part zoning:** From the headline-end position, column-wise consecutive object pixels count are found. The position where the maximum value of that count is attained is noted. If that maximum value becomes compatible with the BB height, then that value is taken as the vertical height of the middle zone of the script (M). U and D, are derived from M and the height of the BB.

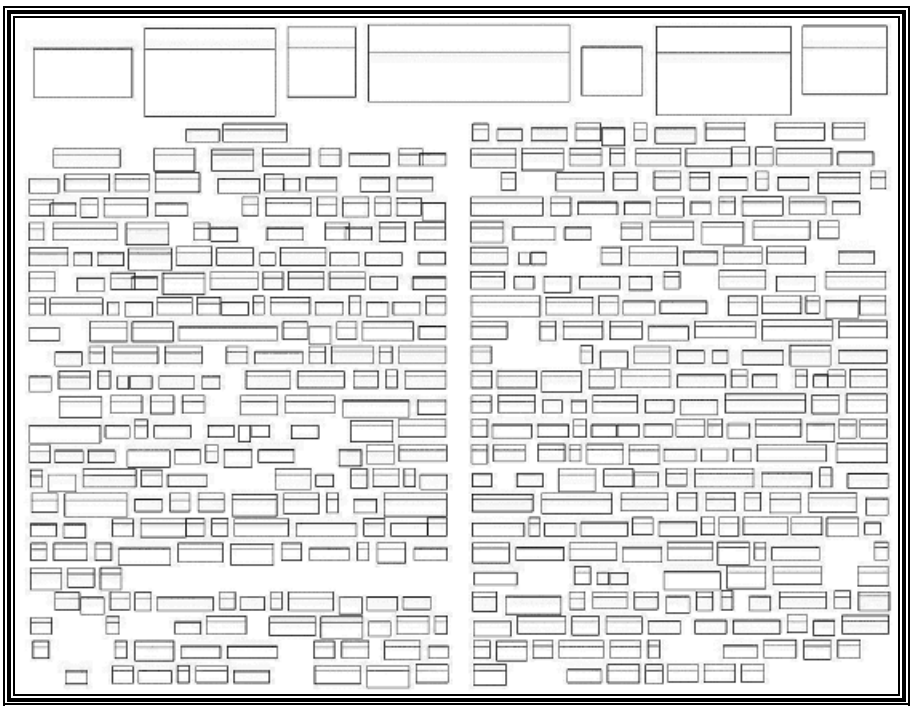


**Figure 3. All bounding boxes with headlines.**

### 4.3. Block construction using primary features (Module I)

Step 1: Consider all BBs with headlines. Initialise block-label by $I$. (at the starting of the process, $I = 1$).

Step 2: Within the width span of the first encountered unlabeled BB with headline say, $B$, find the closest BB, if any, which is also compatible with $B$. If that BB is $B'$, then the vertical headline-to-headline distance between $B$ and $B'$ is computed. Suppose that is $VHH(B, B')$.

Step 3: Similarly, find the closest BB say $B''$ of $B'$, which is also height-wise compatible.

Step 4: If $VHH(B',B'')$ is compatible with $VHH(B,B')$ then put the same block-label to $B$, $B'$, and $B''$. Else, skip B and go to step 2.

Step 5: All other immediate and compatible adjacent (lower and upper) BBs having compatible headline-to-headline distance will be labelled by the block-label $I$ successively.

Step 6: When no more BB is found which is compatible with a BB with block-label $I$ then block-wise feature vectors for all BBs with block $I$ are computed.

Step 7: Repeat from Step 2 by incrementing $I$ by one until each BB gets a block-label number.

Considering the size, number of BBs, existence of headlines etc., all generated blocks and remaining BBs are labelled into seven categories. A block is marked with $C_1$ if it has several (subject to a threshold) compatible BBs or with $C_2$, if there are only a few compatible BBs. $C_3$ is used, if the block has only a single BB with a headline and with a moderate height (in terms of the BBs in its neighbouring BBs). Next, BBs with no headline are considered and are labelled by $C_4$, $C_5$ or $C_6$ respectively, if the size is small, moderate, or large (all these relative measures are computed in terms of the average BB size of the closest block of type $C_1$). Among $C_5$ and $C_6$ labelled BBs, any BB where height-width ratio is highly disproportionate or density is very low, is marked by $C_7$. After processing Module I, all blocks with label $C_1$ are assigned as textual blocks and their union identifies the main textual body of the document. All BBs in blocks with label $C_2$ or $C_3$ are also part of some textual block. However, verification and appropriate modifications may require for finding the actual boundary of these blocks. Note that in all single-line textual blocks such as titles, headers, captions etc., growing was not possible at the Module I. So all words and characters without headlines are separated. Most of BBs of type $C_3$ reside in these blocks. On the other hand, if in a regular paragraph

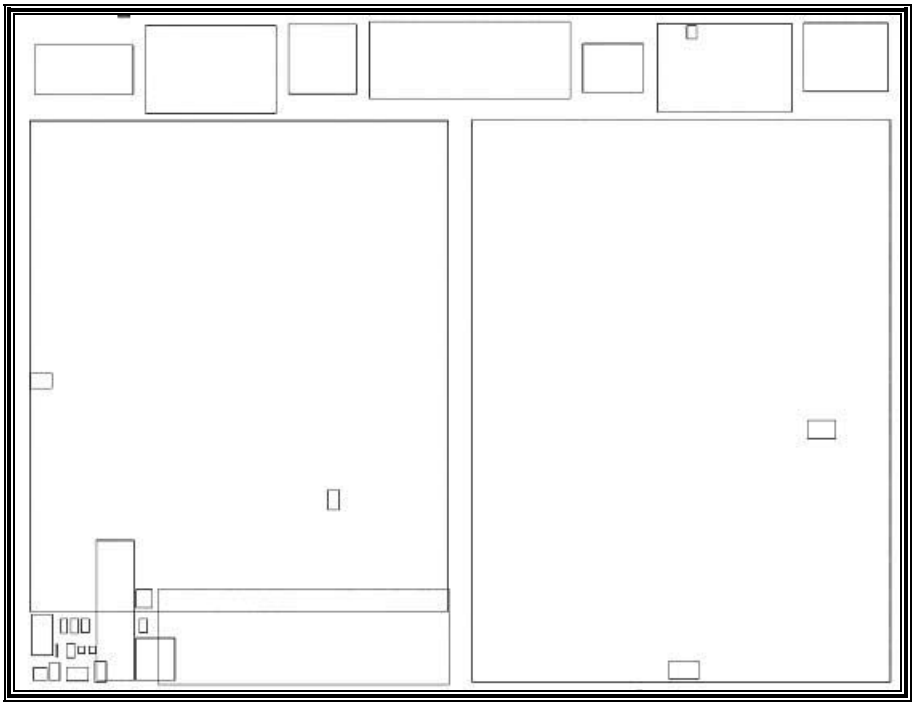


**Figure 4. Total blocks after initial grouping.**

(typically of only a few lines height) there is a space at the same horizontal position on all lines, then an inconsistent block with type $C_2$ or $C_3$ may appear at that location. Any BB with label $C_4$ is likely to be a part of a character only if it is sufficiently close to another BB, already assigned to a textual block. On the other hand, any BB with label $C_5$ is likely to be a part of some textual block only if it is sufficiently close to another BB within a textual block and they are height-wise compatible.

In the following, we briefly discuss the modules used for block type rectification and correction of block boundaries. These modules are repeated twice one after another.

## 4.4. Block absorption (Module II)

Let there be a block (say,) $F_{in}$ completely inside another block (say,) $F_{out}$ of type other than $C_7$. Let $F_{in}$, of type $C_2$, be compatible with $F_{out}$ in terms of vertical heights of the middle zone M. If the pair-wise closest horizontal distance between horizontally aligned (having common projection in their heights) BBs of $F_{in}$ and $F_{out}$ is compatible with the average horizontal gap of BBs within $F_{out}$, then $F_{in}$ is absorbed in $F_{out}$. Labels of the BBs in $F_{in}$ are updated by the block label of $F_{out}$. This type of assimilation removes all inconsistent blocks due to spacing at the same horizontal position on all lines within a single paragraph. If $F_{in}$ is of any type other than $C_1$, $C_2$ and $C_7$, the compatibility criterion given above is used after increasing the value of compatibility parameter ε. As a result, all BBs without headlines but within a paragraph already being identified by a textual block are absorbed. Particularly, all split-characters and those with no headline and separated from words are taken care of. Now, consider the case where $F_{in}$ is partially within $F_{out}$. If $F_{out}$ is considerably bigger than $F_{in}$ and the average height of the middle zone of $F_{in}$ is compatible but lesser than that of $F_{out}$, $F_{in}$ is merged into $F_{out}$. However, since loose compatibility criterion is applied, for the last two cases, all BBs in these $F_{in}$s are marked with doubt keys. This is useful to identifying enlarged starting character of a document, which often occurs in media publications. (See Figure 6.)

## 4.5. Block merging (Module III)

The blocks, not labelled by $C_1$ are considered. If some of them are spatially clustered then those blocks are merged and a bigger block is generated. Clusters are formed in the following manner. Among these blocks, let $F$ be the closest horizontally aligned block of $F'$ at a horizontal distance $H$ apart. Also, let $H$ be compatible with the average horizontal distances between adjacent BBs within each block, if exist. Now, if average height

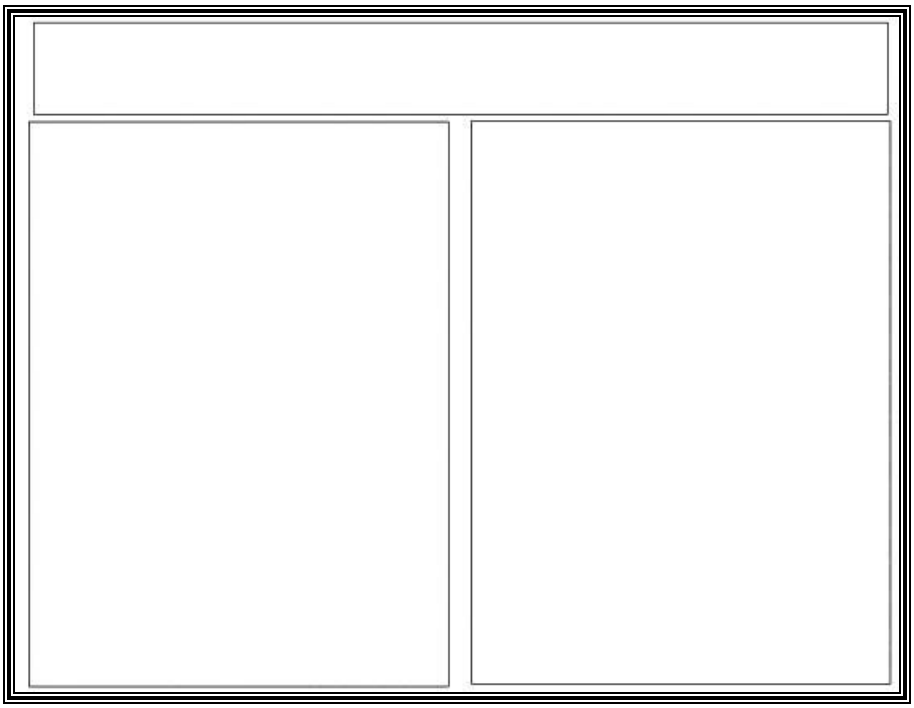


**Figure 5. Finally derived three textual blocks.**

and density of these two blocks are loosely compatible then those blocks are merged.

## 5. Experimental Results

During testing period, we use a large number of document images of both Bangla and Devnagari pages with multi-column layout and some with small sized figures as well. Altogether hundred pages are used for testing. Forty pages are from '*Desh*', the most popular Bangla literature magazine and '*The Anaandabazar Patrika*' which has the maximum readership as a daily newspaper. Thirty pages are derived from the '*Sanmarga Patrika*', one of the most popular newspapers in Hindi. Five pages for each script, from popular books and other newspapers, magazines etc., ten pages for each script, from hard copies of computer generated documents with various font sizes and shapes, are taken. These pages are scanned in 300 dpi. The page sizes are different for different documents and on an average have 2500 x 2000 pixels. In Figures 2-5, the input and output of the page-layout analyser at the different stages of the algorithm are shown on a Hindi document.

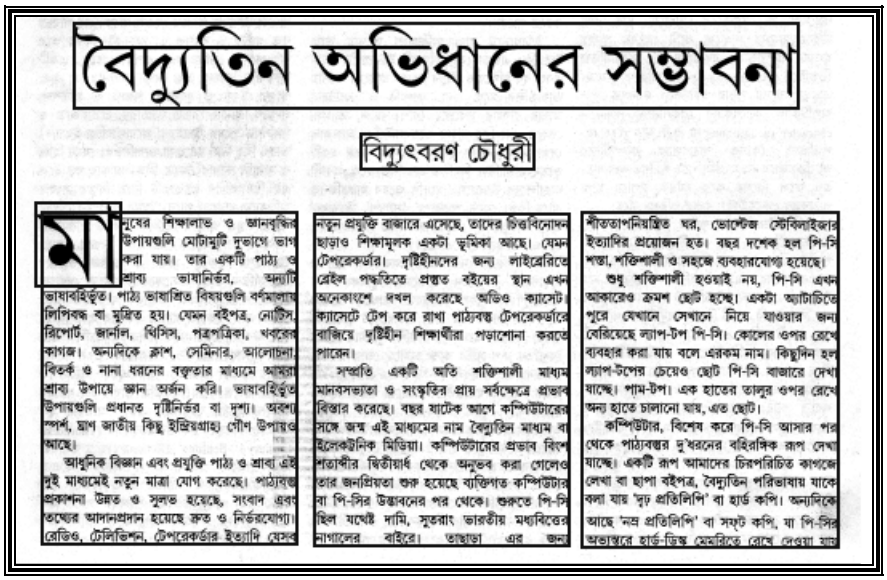
The pages are correctly segmented into textual regions in about 98% cases. Wrong decomposition occurs mostly in the periphery of a text-region, such as page numbering zone where all is covered by few numerals or scripts in a language such as English having any headline.

Figure 6 and Figure 7 show two more results for Bangla documents. Note that in Figure 7, the image zone and the rulers are correctly separated from the detected text blocks.

The system is implemented in a Pentium, III, 700 MHz machine using Visual C++. The computation is considerably fast. It takes on an average 1.35 seconds for a complete page layout segmentation. The average time includes the processing times for all modules starting from reading the TIFF file to final block-level grouping.

## 6. Summary and Discussion

In this paper, the present status of the ongoing project on page layout analysis of Indian documents is reported. The analyser uses the common textual features present in most of the characters in Indian scripts derived from Brahmi, such as headlines in the upper zone, vertical lines in the middle zone and word level connectivity. The initial block-level grouping is performed based on homogeneity among the vertical adjacent bounding boxes. Correction of this initial grouping is also performed. The results are restricted on images of two most popular Indian scripts viz., Devnagari and Bangla. The choice is



বৈদ্যুতিন অভিধানের সম্ভাবনা

বিদ্যুৎবরণ চৌধুরী

মানুষের শিক্ষালাভ ও জ্ঞানবৃদ্ধির উপায়গুলি মোটামুটি দুভাগে ভাগ করা যায়। তার একটি পাঠ্য ও শ্রাব্য ভাষানির্ভর, অন্যটি ভাষাবহির্ভূত। পাঠ্য ভাষাশ্রিত বিষয়গুলি বর্ণমালায় লিপিবদ্ধ বা মুদ্রিত হয়। যেমন বইপত্র, নোটিস, রিপোর্ট, জার্নাল, থিসিস, পত্রপত্রিকা, খবরের কাগজ। অন্যদিকে ক্লাশ, সেমিনার, আলোচনা, বিতর্ক ও নানা ধরনের বক্তৃতার মাধ্যমে আমরা শ্রাব্য উপায়ে জ্ঞান অর্জন করি। ভাষাবহির্ভূত উপায়গুলি প্রধানত দৃষ্টিনির্ভর বা দৃশ্য। অবশ্য স্পর্শ, ঘ্রাণ জাতীয় কিছু ইন্দ্রিয়গ্রাহ্য গৌণ উপায়ও আছে।

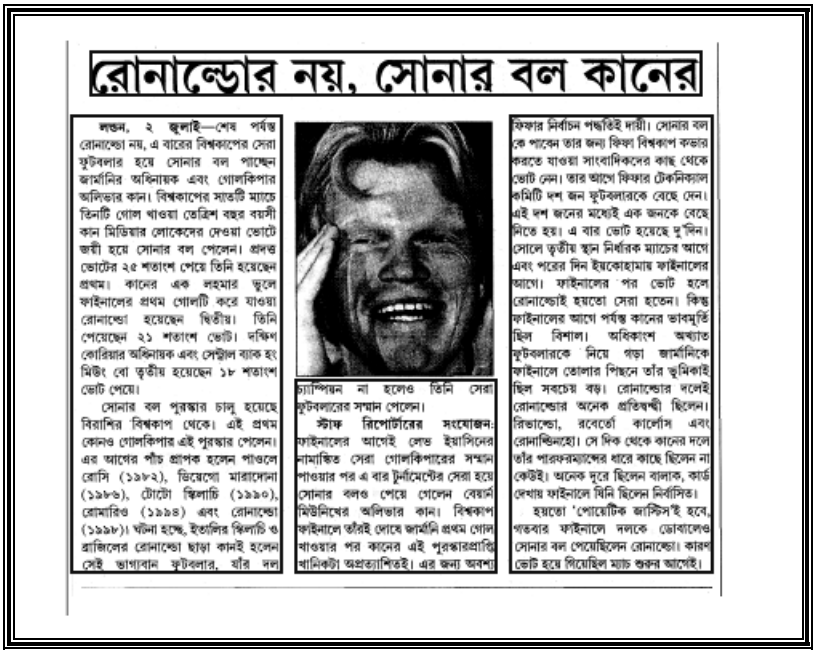
আধুনিক বিজ্ঞান এবং প্রযুক্তি পাঠ্য ও শ্রাব্য এই দুই মাধ্যমেই নতুন মাত্রা যোগ করেছে। পাঠ্যবস্তু প্রকাশনা উন্নত ও সুলভ হয়েছে, সংবাদ এবং তথ্যের আদানপ্রদান হয়েছে দ্রুত ও নির্ভরযোগ্য। রেডিও, টেলিভিশন, টেপরেকর্ডার ইত্যাদি যেসব নতুন প্রযুক্তি বাজারে এসেছে, তাদের চিত্তবিনোদন ছাড়াও শিক্ষামূলক একটা ভূমিকা আছে। যেমন টেপরেকর্ডার। দৃষ্টিহীনদের জন্য লাইব্রেরিতে ব্রেইল পদ্ধতিতে প্রস্তুত বইয়ের স্থান এখন অনেকাংশে দখল করেছে অডিও ক্যাসেট। ক্যাসেটে টেপ করে রাখা পাঠ্যবস্তু টেপরেকর্ডারে বাজিয়ে দৃষ্টিহীন শিক্ষার্থীরা পড়াশোনা করতে পারেন।

সম্প্রতি একটি অতি শক্তিশালী মাধ্যম মানবসভ্যতা ও সংস্কৃতির প্রায় সর্বক্ষেত্রে প্রভাব বিস্তার করেছে। বছর ষাটেক আগে কম্পিউটারের সঙ্গে জন্ম এই মাধ্যমের নাম বৈদ্যুতিন মাধ্যম বা ইলেকট্রনিক মিডিয়া। কম্পিউটারের প্রভাব বিংশ শতাব্দীর দ্বিতীয়ার্ধ থেকে অনুভব করা গেলেও তার জনপ্রিয়তা শুরু হয়েছে ব্যক্তিগত কম্পিউটার বা পি-সির উদ্ভাবনের পর থেকে। শুরুতে পি-সি ছিল যথেষ্ট দামি, সুতরাং ভারতীয় মধ্যবিত্তের নাগালের বাইরে। তাছাড়া এর জন্য শীতাতপনিয়ন্ত্রিত ঘর, ভোল্টেজ স্টেবিলাইজার ইত্যাদির প্রয়োজন হত। বছর দশেক হল পি-সি শস্তা, শক্তিশালী ও সহজে ব্যবহারযোগ্য হয়েছে।

শুধু শক্তিশালী হওয়াই নয়, পি-সি এখন আকারেও ক্রমশ ছোট হচ্ছে। একটা অ্যাটাচিতে পুরে যেখানে সেখানে নিয়ে যাওয়ার জন্য বেরিয়েছে ল্যাপ-টপ পি-সি। কোলের ওপর রেখে ব্যবহার করা যায় বলে এরকম নাম। কিছুদিন হল ল্যাপ-টপের চেয়েও ছোট পি-সি বাজারে দেখা যাচ্ছে। পাম-টপ। এক হাতের তালুর ওপর রেখে অন্য হাতে চালানো যায়, এত ছোট।

কম্পিউটার, বিশেষ করে পি-সি আসার পর থেকে পাঠ্যবস্তুর দু'ধরনের বহিরঙ্গিক রূপ দেখা যাচ্ছে। একটি রূপ আমাদের চিরপরিচিত কাগজে লেখা বা ছাপা বইপত্র, বৈদ্যুতিন পরিভাষায় যাকে বলা যায় 'দৃঢ় প্রতিলিপি' বা হার্ড কপি। অন্যদিকে আছে 'মৃদু প্রতিলিপি' বা সফ্‌ট কপি, যা পি-সির অভ্যন্তরে হার্ড-ডিস্ক মেমরিতে রেখে দেওয়া যায়

**Figure 6. Page-layout result on a Bangla document taken from the magazine *Desh*.**



রোনাল্ডোর নয়, সোনার বল কানের

লন্ডন, ২ জুলাই—শেষ পর্যন্ত রোনাল্ডো নয়, এ বারের বিশ্বকাপের সেরা ফুটবলার হয়ে সোনার বল পাচ্ছেন জার্মানির অধিনায়ক এবং গোলকিপার অলিভার কান। বিশ্বকাপের সাতটি ম্যাচে তিনটি গোল খাওয়া তেত্রিশ বছর বয়সী কান মিডিয়ার লোকেদের দেওয়া ভোটে জয়ী হয়ে সোনার বল পেলেন। প্রদত্ত ভোটের ২৫ শতাংশ পেয়ে তিনি হয়েছেন প্রথম। কানের এক লহমার ভুলে ফাইনালের প্রথম গোলটি করে যাওয়া রোনাল্ডো হয়েছেন দ্বিতীয়। তিনি পেয়েছেন ২১ শতাংশ ভোট। দক্ষিণ কোরিয়ার অধিনায়ক এবং সেন্টার ব্যাক হং মিউং বো তৃতীয় হয়েছেন ১৮ শতাংশ ভোট পেয়ে।

সোনার বল পুরস্কার চালু হয়েছে বিরাশির বিশ্বকাপ থেকে। এই প্রথম কোনও গোলকিপার এই পুরস্কার পেলেন। এর আগের পাঁচ গ্রাপক হলেন পাওলো রোসি (১৯৮২), ডিয়েগো মারাদোনা (১৯৮৬), টোটো স্কিলাচি (১৯৯০), রোমারিও (১৯৯৪) এবং রোনাল্ডো (১৯৯৮)। ঘটনা হচ্ছে, ইতালির স্কিলাচি ও ব্রাজিলের রোনাল্ডো ছাড়া কানই হলেন সেই ভাগ্যবান ফুটবলার, যাঁর দল



চ্যাম্পিয়ন না হলেও তিনি সেরা ফুটবলারের সম্মান পেলেন।

**স্টাফ রিপোর্টারের সংযোজন:** ফাইনালের আগেই লেভ ইয়াসিনের নামাঙ্কিত সেরা গোলকিপারের সম্মান পাওয়ার পর এ বার টুর্নামেন্টের সেরা হয়ে সোনার বলও পেয়ে গেলেন বেয়ার্ন মিউনিখের অলিভার কান। বিশ্বকাপ ফাইনালে তাঁরই দোষে জার্মানি প্রথম গোল খাওয়ার পর কানের এই পুরস্কারপ্রাপ্তি খানিকটা অপ্রত্যাশিতই। এর জন্য অবশ্য ফিফার নির্বাচন পদ্ধতিই দায়ী। সোনার বল কে পাবেন তার জন্য ফিফা বিশ্বকাপ কভার করতে যাওয়া সাংবাদিকদের কাছ থেকে ভোট নেন। তার আগে ফিফার টেকনিক্যাল কমিটি দশ জন ফুটবলারকে বেছে দেন। এই দশ জনের মধ্যেই এক জনকে বেছে নিতে হয়। এ বার ভোট হয়েছে দু'দিন। সোলে তৃতীয় স্থান নির্ধারক ম্যাচের আগে এবং পরের দিন ইয়কোহামায় ফাইনালের আগে। ফাইনালের পর ভোট হলে রোনাল্ডোই হয়তো সেরা হতেন। কিন্তু ফাইনালের আগে পর্যন্ত কানের ভাবমূর্তি ছিল বিশাল। অধিকাংশ অখ্যাত ফুটবলারকে নিয়ে গড়া জার্মানিকে ফাইনালে তোলার পিছনে তাঁর ভূমিকাই ছিল সবচেয়ে বড়। রোনাল্ডোর দলেই রোনাল্ডোর অনেক প্রতিদ্বন্দ্বী ছিলেন। রিভাল্ডো, রবের্তো কার্লোস এবং রোনাল্ডিনহো। সে দিক থেকে কানের দলে তাঁর পারফরম্যান্সের ধারে কাছে ছিলেন না কেউই। অনেক দূরে ছিলেন বালাক, কার্ড দেখায় ফাইনালে যিনি ছিলেন নির্বাসিত।

হয়তো 'পোয়েটিক জাস্টিস'ই হবে, গতবার ফাইনালে দলকে ডোবালেও সোনার বল পেয়েছিলেন রোনাল্ডো। কারণ ভোট হয়ে গিয়েছিল ম্যাচ শুরুর আগেই।

**Figure 7. Page-layout result on a Bangla document taken from the newspaper *Anaandabazar*.**

dictated by (1) the availability of large number of documents of this type, (2) existence of basic OCR modules on these languages that work only for single column plain text documents and (3) absence of any work on page layout analysis dedicated for Indian scripts.

Note that at present, no pre-processing is included for image zone and other non-textual zone localization before textual region identification. We are devising some modules that could identify and categorize these regions. We are also going to incorporate a module that could successfully handle table like structures.

## 7. References


[1] Nagy, G, 'Twenty Years of Document Image Analysis', *IEEE Trans. Pattern Analysis and Machine Intelligence,* 22/1, 2000, pp. 38-62.

[2] Cottoni, R. Coianiz, T. Messelodi, S. Modena, C.M., 'Geometric Layout Analysis Technique for Document Image Understanding: A Review', *ITC-irst TR* #9703-09, 1997.

[3] Johnston, E.G.,'SHORT NOTE: Printed Text Discrimination', *Computer Graphics and Image Processing,* Vol. 3, 1974, pp.83-89.

[4] O'Gorman, L, 'The Document Spectrum for Page Layout Analysis', *IEEE Trans. Pattern Analysis and Macine Intelligence,* 15/11, 1993, pp.1162-1173.

[5] Jain, A. K. and Yu. B, 'Document Representation and Its Application to Page Decomposition', *IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20,*3, 1998, pp. 294-308.

[6] Srihari , S.N and Govindaraju, V, 'Analysis of Textual Images Using the Hough Transform', *Machine Vision & Application, 2/3,* 1989, pp. 141-153.

[7] Pavlidis., T. and Zhou, J., 'Page Segmentation and Classification', *Graphical Models and Image Processing, 54/6,* 1992, pp. 484-496.

[8] Chen. S., Haralick, R.M. and Phillips I.P, 'Extraction of text layout structure on Document Images based on Statistical Characterization', *ISET/SPIE Symposium on Electronic Imaging Science & Technology Document Recognition II,* 1995, pp. 128-139.

[9] Baird, H.S, 'Background Structure in Document Images.' *In Advance in Structural and Symmetric Pattern recognition, World Scientific,* 1992, pp. 253-269.

[10] Chaudhuri, B.B. and Pal, U, 'Relational Studies between Phoneme and Grapheme Statistics in Modern Bangla Language', *J.Acoustical Society of India, 23,* 1995, pp. 67-77.