International Journal of Scientific Research in Computer Science and Engineering **Research Paper**
Vol.6, Issue.2, pp.51-56, April (2018) **E-ISSN:** 2320-7639

XY Cut Modular approach for Segmenting pages

Simple Batra

IT Dept, Delhi School of Professional Studies and Research (DSPSR), Guru Gobind Singh Indraprastha University, Dwarka, India. New Delhi, India

E-mail: simpletondon@gmail.com Tel: 7290988620

Available online at: www.isroset.org

Received: 27/Feb/2018, Revised: 12/Mar/2018, Accepted: 31/Mar/2018, Online: 30/Apr/ 2018

***Abstract*—** the purpose of this experimental research is to present algorithm for reading contents of documented image. Most of the information that is available today in the world is in printed medium. Printed data has hindered storing, exchanging and processing of this information electronically. Converting them from printed to electronic medium is time consuming as well as expensive as well. These factors have motivated people to develop automated systems to perform such task. Optical Character Recognition (OCR) is an important technique which is very popular among research and technical communities. As a result of these research and development activities several OCR applications are being made available in the market. In this paper, we propose two separate modules to determining the paragraphs and lines in of a document page which is independent of languages.

***Keywords*—** OCR, document image, X-Y cut segmentation, over segmentation, under segmentation.

I. INTRODUCTION

Optical Character Recognition (OCR) is an automated process of translating an input document image into a symbolic text file. The input document images can come from a large variety of media, such as journals, news- papers, magazines, memos, etc. The format of any document image can be digitally created, faxed, scanned, machine printed, or handwritten. The output symbolic text file from an OCR system can include not only the text content of the input document image but also additional descriptive information, such as page layout, font size and style, document region type and confidence level for the recognized characters.

A complete OCR system includes a full suite of processes from skew correction, binarization, segmentation; text and non-text block classification, line, word and character segmentation and character recognition to final reconstruction. For an optical character recognition (OCR) system, segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation phase. Incorrect segmentation leads to incorrect recognition. Document segmentation is recognized as a hard problem and it may not be possible to formulate a single algorithm, which works for all kinds of documents. Segmentation phase includes block, line, and word and character segmentation. Before word and character segmentation, line segmentation is performed to find the number of lines and boundaries of each line in any input document image. Incorrect segmentation may result in decrease in recognition accuracy. Text extraction plays a significant role in document retrieval and storage systems. In this paper recursive cuts based on top down approach have been implemented using MATLAB. The entire input document page is the root node and the resulted blocks of final segmentations are the leaf nodes. Module 1 segments all paragraphs and Module 2 segments all lines from the document image.

II . OBJECTIVE

The aim is to identify and implement the improved mechanism to segment the pages scanned by an OCR as compared to the existing algorithms used to segment the pages to enhance the efficiency and accuracy level of the system. It aims to develop an algorithm, which can adapt to various document styles and can be robust in the presence of noise and skew. Segment the document into separate regions like blocks, lines, and words

- Separate the blocks using the X-Y cut algorithm.
- Separate the lines from the blocks.

III. LITERATURE SURVEY

There are two types of page segmentation mechanisms, physical and logical. Physical page segmentation is a process of dividing a document page into homogeneous zones. Each of these zones can contain one type of object. These objects can be of any type text, table, a figure, a halftone image etc. Logical page segmentation is a process of assigning logical relations to physical zones. For example, reading order labels order the physical zones in the order in which they should be read. In the same way assigning section and sub section labels to physical zones creates a hierarchical document structure. Page segmentation is a very crucial preprocessing step for any OCR system. In different cases, OCR engine recognition accuracy depends heavily on page segmentation algorithm's accuracy. For instance, if a page segmentation algorithm merges two text zones horizontally then the OCR engine will recognize text across text zones and hence generate unreadable text.

Document Image Layout analysis-

- Document image layout has a hierarchical (tree like) representation, with each level encapsulating some unique information that is not present in other levels.
- The representation contains the complete page in the root node, and the text blocks, images and background form the next layer of the hierarchy.
- The text blocks can have further detailed representations like text lines, words and components, which form the remaining layers of the representation hierarchy.
- Document image layout analysis is a crucial step in many applications related to document images, like text extraction using optical character recognition (OCR), reflowing documents, and layout-based document retrieval.
- Layout analysis is the process of identifying layout structures by analyzing page images.
- Layout structures can be physical (text, graphics, pictures . . .) or logical (titles, paragraphs, captions, headings . . .).
- The identification of physical layout structures is called physical or geometric layout analysis, while assigning different logical roles to the detected regions is termed as logical layout analysis.
- The geometric page layout of a document is to specify the geometry of the maximal homogeneous regions and their classifications (text, table, images, drawing, etc.).The task of a geometric layout analysis system is to segment the document image into homogeneous zones, each consisting of only one physical layout structure, and to identify their spatial relationship (e.g. reading order).
- Logical page layout analysis is to determine the type of page, assign functional labels such as title, logo, footnote, caption, etc., to each block of the page, determine the relationships of these blocks and order the text blocks according to their reading order as a one dimensional stream a linear search reveals the document regions at each scale levels in a nested fashion as text block, text lines and graphics.

Techniques for page segmentation and layout analysis-

Texture based and Non texture-based-.

- The texture-based methods consider a document image as a composite of textures of different classes. With this approach, various well-known texture segmentation techniques can be used but should be modified correspondingly to fit to special cases.
- While the non-texture-based methods apply other image processing techniques (for example, to use various splitting-and-merging strategies) to this task and do not treat the image as texture.

Strategy based-

- Page segmentation algorithms can be categorized into three classes: top-down approaches (or model-driven), bottom-up approaches (or data-driven), hybrid approaches.
- 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. In top-down methods, a document is segmented from large components (high-level) to smaller, more detailed, sub-components (lower-level). It starts by segmenting the document into large blocks and then analyses them in order to achieve separation of the characters in the text blocks. Top-down algorithms start from the whole document image and iteratively split it into smaller ranges. The splitting procedure stops when some criterion is met and the obtained ranges constitute the final segmentation results. The X-Y cut by Nagy et al and the shape-directed-covers-based algorithm by Baird et al. are top-down algorithms.

- 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. The bottom-up method, starts by first segmenting the document into small blocks such as characters, and then merges them into bigger blocks as words and then text lines. Bottom-up algorithms start from document image pixels and cluster the pixels into connected components which are then clustered into words, lines, or final zone segmentations. The Docstrum algorithm of O'Gorman, the Voronoidiagram- based algorithm of Kise et al., the run-length smearing algorithm of Wahl et al., the segmentation algorithm of Jain and Yu , and the text string separation algorithm of Fletcher and Kasturi are typical bottom-up algorithms.
- To some extent, hybrid algorithms are a mix of the above two approaches. Pavlidis and Zhou proposed a hybrid algorithm using a split-and-merge strategy.

IV. PROPOSED APPROACH

XY-Cut for Segmentation of Page Image-

The method consists in finding the widest empty rectangle(s), or valley(s) entirely crossing the page either vertically or horizontally. The page is then segmented in blocks, which are shrunk to fit their content closely. The method applies again recursively to each block. It stops when no sufficiently large valley can be found in any of the created blocks, which become the final image segments.

XY-Cut for Text Ordering-

The XY-cut algorithm induces an order among the blocks generated in hierarchy, in particular among leaf blocks. Typically, for a western conventional reading order, ordering the block hierarchy of block in a top-to bottom left-to-right way leads to the expected order, as illustrated by figure below.

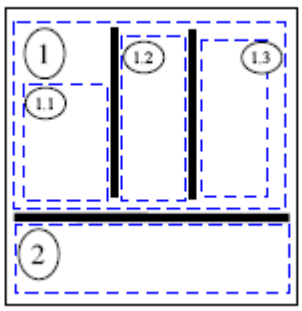


Figure3- Ordering of blocks

Proposed approach-

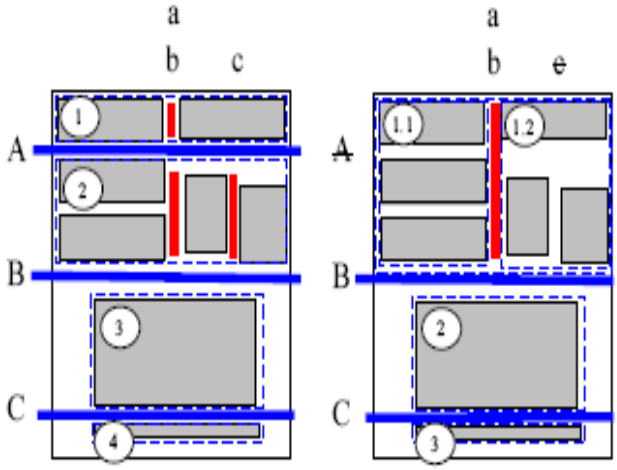


Figure4- Example of all the possible horizontal and vertical cuts defining blocks of a page.

The left side three possible horizontal cuts (A. B, C) defining four possible blocks (1, 2, 3, and 4). Within each block, the possible vertical cuts are shown in red (a, b, c). Seven choices are offered here: taking only one of the horizontal cuts, two of them or all of them. The right side shows the preferred cuts (B, C to make a + b possible for the first block).

In more detail, the block cutting process becomes:

1. Given a block to segment, the method enumerates all possible horizontal cuts; (A, B and C on the previous figure)

2. For each block potentially created by an enumerated horizontal cut, the method enumerates all possible vertical cuts inside it; ((a) (b, c) (e))

3. A (possibly empty) set of horizontal cuts is chosen so as to permit the best possible series of enumerated vertical cuts, given the score function.

4. The selected horizontal cuts are performed, and then for each created block the set of associated vertical cuts is performed as well.

5. Return to step 1 for each created block until no further cuts are possible.

Module 1-

PAGE SEGMENTATION-

Module 1 works on document images and does the segmentation on the block level. This module has a recursive function named as paragraph ( ) which gives the optimal cuts in the input document image given to it in order to separate

the paragraphs from it. On the basis of the intensity of the pixels the breakage points are defined for cuts. Threshold value gives the optimum value for the cut to be enabled as per the document.

Steps used in algorithm-

1. Reading the Document image.
2. Converting it into binary image.
3. Defining the matrix of its size.
4. Defining the breakage points.
5. Setting the threshold value for the cuts.
6. Within the loop with every cut the bounding box is defined around each paragraph.

Module 2-

LINE SEGMENTATION

Module 2 takes a paragraph or any document image as an input and a function named as "lineext" is called, which works recursively to extract the lines from that document. Line segmentation is an operation that belongs to the page segmentation class of problems. Although grey level images can be used, in general the segmentation is performed on binary version of the images. This will improve the recognition rate since that helps the algorithm to differentiate the background and the foreground by inspecting the pixel intensity. The threshold for corresponding binary values is determined dynamically by inspecting the most frequent and least frequent pixel intensities. When a binary values are available for a document, extraction of the text lines can be done by a connected component collection and grouping. We first generate the connected component representation based on the pixel intensities. For each text line pattern, collect all the connected components touching the pattern, these components together make up the text line. The space between text lines should be identified and examined. In a page image, text lines belonging to the same paragraph are vertically spaced apart by practically the same inter-line distance. Text line segmentation is a labeling process which consists in assigning the same label to spatially aligned units (such as pixels, connected components or characteristic points). There are two categories of text line segmentation approaches: searching for (fictitious) separating lines or paths, or searching for aligned physical units. The choice of a segmentation technique depends on the complexity of the text line structure of the document. The RXY cuts method applied uses alternating projections along the X and the Y axis. This results in a hierarchical tree structure. Cuts are found within whitespaces. Thresholds are necessary to derive inter-line or inter-block distances. This method can be applied to printed documents (which are assumed to have these regular distances).

Steps used in Module 2 –

1. Takes a paragraph in the form of document image.
2. Finds the binary values of the image to inspect the varying intensities in the image.
3. Defines the matrix of its size.
4. Stores the pixel values in the cells of matrix.
5. Checks the intensity of pixels of the one row if found as a white pixel means with value 0 continues till it is found as 0.
6. It takes a break when the black pixel is found and jumps to the next line and checks again if the first pixel of the next row is also a white pixel.
7. It recursively gives cuts within the whitespaces.
8. The lines so obtained are segmented regions around stored in bounding boxes in same file.

V. RESULT

To implement XY Cut page segmentation any scanned original document is given as an input and two modules are implemented on original document using MATLAB.

Result of module 1

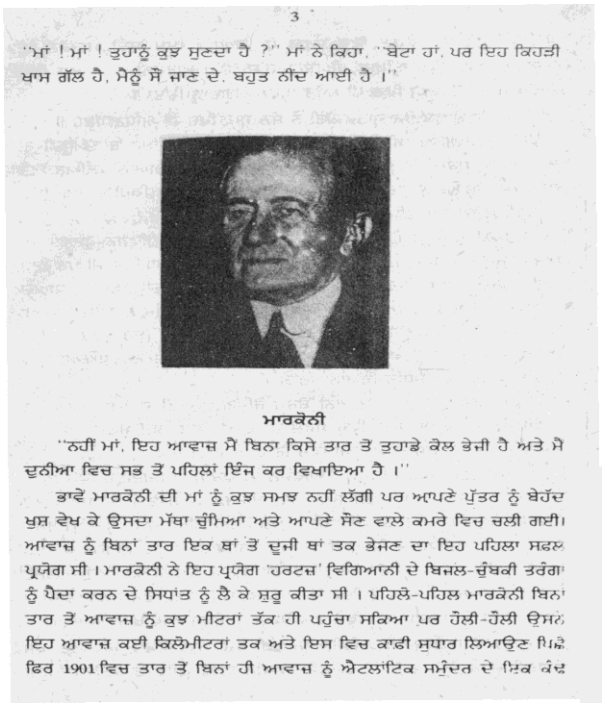


Figure 5 this is the input image given to the module1 as an input for the extraction of the blocks from it

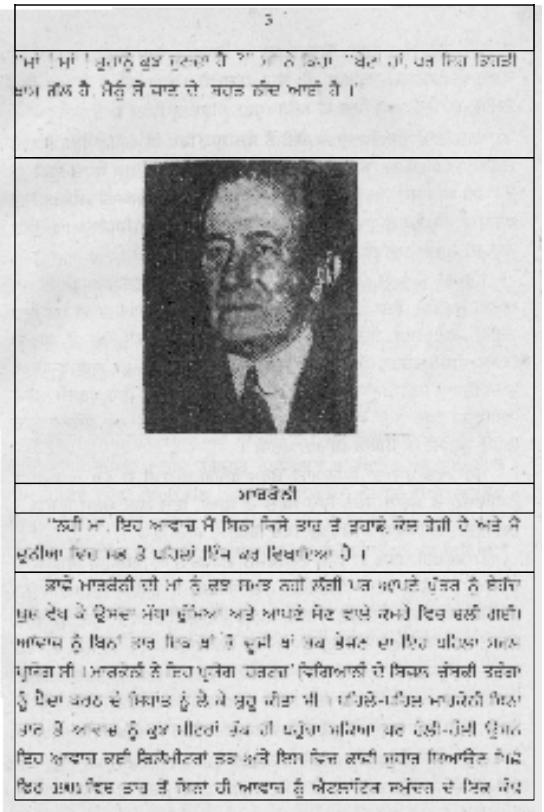
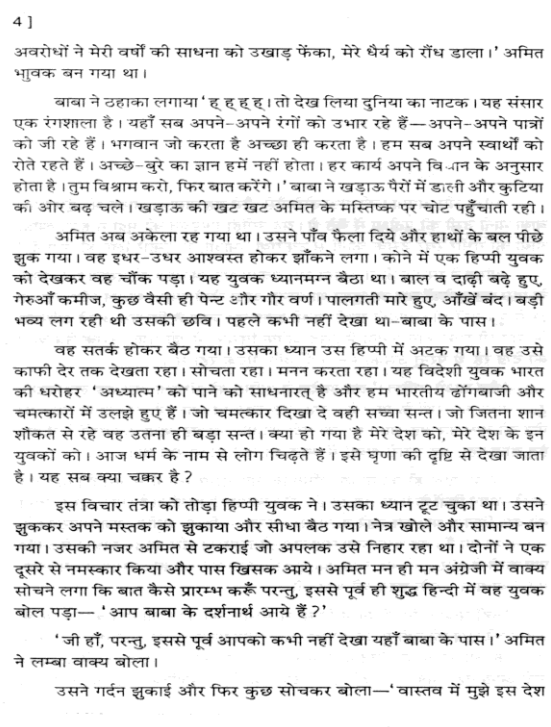


Figure6- Showing all the paragraphs obtained after implementation of module 1 in Input Image 1



4]

अवरोधों ने मेरी वर्षों की साधना को उखाड़ फेंका, मेरे धैर्य को रौंध डाला।' अमित भावुक बन गया था।

बाबा ने ठहाका लगाया 'ह ह ह ह। तो देख लिया दुनिया का नाटक। यह संसार एक रंगशाला है। यहाँ सब अपने-अपने रंगों को उभार रहे हैं—अपने-अपने पात्रों को जी रहे हैं। भगवान जो करता है अच्छा ही करता है। हम सब अपने स्वार्थों को रोते रहते हैं। अच्छे-बुरे का ज्ञान हमें नहीं होता। हर कार्य अपने विधान के अनुसार होता है। तुम विश्राम करो, फिर बात करेंगे।' बाबा ने खड़ाऊ पैरों में डाली और कुटिया की ओर बढ़ चले। खड़ाऊ की खट खट अमित के मस्तिष्क पर चोट पहुँचाती रही।

अमित अब अकेला रह गया था। उसने पाँव फैला दिये और हाथों के बल पीछे झुक गया। वह इधर-उधर आश्वस्त होकर झाँकने लगा। कोने में एक हिप्पी युवक को देखकर वह चौंक पड़ा। यह युवक ध्यानमग्न बैठा था। बाल व दाढ़ी बढ़े हुए, गेरुआँ कमीज, कुछ वैसी ही पेन्ट और गौर वर्ण। पालगती मारे हुए, आँखें बंद। बड़ी भव्य लग रही थी उसकी छवि। पहले कभी नहीं देखा था-बाबा के पास।

वह सतर्क होकर बैठ गया। उसका ध्यान उस हिप्पी में अटक गया। वह उसे काफी देर तक देखता रहा। सोचता रहा। मनन करता रहा। यह विदेशी युवक भारत की धरोहर 'अध्यात्म' को पाने को साधनारत् है और हम भारतीय ढोंगबाजी और चमत्कारों में उलझे हुए हैं। जो चमत्कार दिखा दे वही सच्चा सन्त। जो जितना शान शौकत से रहे वह उतना ही बड़ा सन्त। क्या हो गया है मेरे देश को, मेरे देश के इन युवकों को। आज धर्म के नाम से लोग चिढ़ते हैं। इसे घृणा की दृष्टि से देखा जाता है। यह सब क्या चक्कर है?

इस विचार तंत्रा को तोड़ा हिप्पी युवक ने। उसका ध्यान टूट चुका था। उसने झुककर अपने मस्तक को झुकाया और सीधा बैठ गया। नेत्र खोले और सामान्य बन गया। उसकी नजर अमित से टकराई जो अपलक उसे निहार रहा था। दोनों ने एक दूसरे से नमस्कार किया और पास खिसक आये। अमित मन ही मन अंग्रेजी में वाक्य सोचने लगा कि बात कैसे प्रारम्भ करूँ परन्तु, इससे पूर्व ही शुद्ध हिन्दी में वह युवक बोल पड़ा— 'आप बाबा के दर्शनार्थ आये हैं?'

'जी हाँ, परन्तु, इससे पूर्व आपको कभी नहीं देखा यहाँ बाबा के पास।' अमित ने लम्बा वाक्य बोला।

उसने गर्दन झुकाई और फिर कुछ सोचकर बोला—'वास्तव में मुझे इस देश

Figure 7-Input Image 2

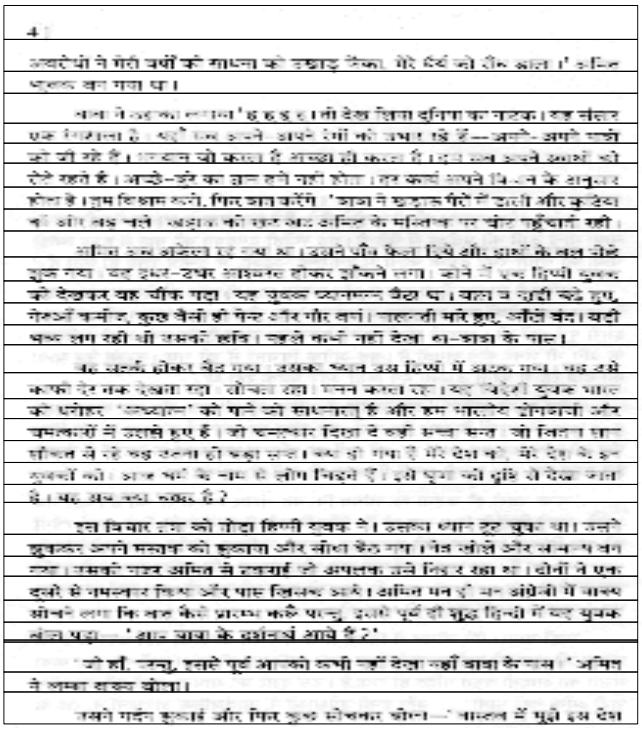


Figure 8-Output of Image

VI. CONCLUSION AND FUTURE WORK

In this paper two modules have been implemented independent of each other and based on the recursive approach to perform the horizontal and vertical projections and cuts are found within the whitespaces based on the varied intensities of the pixels. Thresholds are necessary to find out the inter-line or inter-block distances. We need to find out the coordinates of the segmented regions in order to plot the bounding box around that segmented regions. As per the varied document styles and depending upon the style of documents it becomes complex to define the optimal cuts in the document. The method can be applied to the printed documents and are independent of language.

In future the work can be done for reduction of border noise, enhancement of the segmentation of paragraphs, lines and words from the multicolumn documents and which include graphics positions at different places in order to enhance the efficiency of the modules for any OCR. Even research work can be done to look out for some another mechanism, which can work on different font styles and thus makes it independent of font styles even.

VII. REFERENCES

[1]. Faisal Shafait, Daniel Keysers, Thomas M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images", IEEE, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) 2006

[2]. Antonacopoulos1, B. Gatos2 and D. Karatzas, "Page Segmentation Competition" Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 2003 IEEE.

[3]. Faisal Shafait, Daniel Keysers, and Thomas M. Breuel, "Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms" DRAFT, November 30, 2007

[4]. Song Mao and Tapas Kanungo, "Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 23, NO. 3, MARCH 2001

[5]. Zhixin Shi and Venu Govindaraju, "Multi-scale Techniques for Document Page Segmentation", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05) 2005 IEEE

[6]. Jean-Luc MeunierXerox Research Centre Europe, « Optimized XY-Cut for Determining a Page Reading Order", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05) 2005 IEEE

