

A Novel Scheme for the Extraction of Textual Areas of a Scanned Document Using Page Layout Segmentation Algorithm

Jyoti¹ Er. Amit Ranjan²

¹Research Scholar ²Assistant Professor

^{1,2}Department of Computer Science & Engineering

^{1,2}OM Institute of Technology and Management Hisar, India

Abstract— Text extraction using page layout segmentation algorithm in a scanned document is a challenging task in the computer vision. This technique plays a very important role in providing useful and valuable information. Text extraction is a major component for document or textural image analysis. There are various factors texts in documents depend upon such as language, styles, font, sizes, color, background, orientation, fluctuating text lines, crossing or touching text lines. The ascending approach and many other methods to segmentation of scanned documents in the area of background, text, and photographs are considered. Such different algorithms can also be used in the printing industry for selective or enhanced scanning and object-oriented rendering. A page-layout-segmentation technique to extract text from scanned documents has proposed.

Key words: Text Extraction, Page Layout Segmentation Algorithm

I. INTRODUCTION

Document image segmentation to text lines and words is a critical stage towards unconstrained handwritten document recognition. Variation of the skew angle between text lines or on identical text line, existence of overlapping or touching lines, variable character size and non-Manhattan layout are the challenges of text line extraction. Due to high variability of writing methods, scripts, etc., ways that don't use any previous information and adapt to the properties of the document image, as the proposed, would be more robust. Line extraction techniques could also be classified as primarily based, grouping, smearing and Hough-based [1].

Global projections based approaches are very effective for machine printed documents but cannot handle text lines with different skew angles. However, they can be applied for skew correction in documents with constant skew angle [2].

Hough-based strategies handle documents with variation within the skew angle between text lines, however aren't terribly effective once the skew of a text line varies on its dimension [3].

With the forceful advancement in engineering & communication technology, the trendy society is getting into to the data edge.

In change in the traditional document system (paper etc), people now follow electronic document system (PDF Format) for communication and storage which is currently imperative. But on complex matters, the document image is difficult to accurately identify the information directly out of the need.

On such cases preprocessing the document is completed before its entry.

Image segmentation theory, as digital image process has become a vital a part of folks active analysis.

Image process document image segmentation theory is a vital analysis topic within the method it's primarily between the document image pre-processing and advanced character recognition a vital link between.

The comparatively effective and usually used for document image segmentation and classification strategies embrace threshold, and geometric analysis and other categories.

After segmenting, Text half is detected and extracted for any method, earlier, text extraction techniques are developed solely on monochrome documents [4].

These techniques can be classified as bottom-up, top-down and hybrid. Later with the increasing need for color documents, techniques [5] have been proposed

The Segmentation subdivides an image into its constituent region or objects. The level to that the subdivision is carried depends on the matter being solved. That is segmentation ought to stop once the item of interest in Associate in Nursing application are isolated. The segmentation of nontrivial pictures is one amongst the foremost tough tasks in image process. Segmentation accuracy determines the ultimate success or failure of processed analysis procedures.

The text character contain in the document image can be any gray scale value, low resolutions, variable size and embedded in complex background many issues encountered within the segmentation, these includes the distinction within the skew angle between lines, characters or maybe on identical text line, adjacent text line, overlapping words and touching characters.

A. Propose

In this paper the segmentation is proposed in three stages:

- Line segmentation in which we identify the line in the documents
- Word segmentation in which we identify the words in the documents
- Character segmentation in which we identify the character in the documents

The goal of the segmentation is to modify or amendment the illustration of the image into one thing that's additional meaning and easier to investigate

In texture-based approach the input image is typically thought of as a composite of text and non-text or text, image and background texture categories.

Many segmentation algorithms use a categorification window of a definite size within the hope that each one or majority of pixels within the window belong to a similar class [6]. Thereafter, a classification algorithmic rule is often accustomed to label every window within the feature area. For example, in [7] the number of classes is two, and a 2-means classification is used to classify each block of the

image as text or non-text according to its local energy in the wavelet transform domain.

II. LITERATURE REVIEW

A. Image Segmentation for Text Extraction

This paper presents a methodology for extracting text from images such as document images, scene images etc. Text that appears in these images contains important and useful information. Text extraction in images has been used in large variety of applications such as mobile robot navigation, document retrieving, object identification, vehicle license plate detection, etc. In this paper, we employ discrete wavelet transform (DWT) for extracting text information from complex images. The input image may be a colour image or a grayscale image. If the image is colour image, then preprocessing is required. For extracting text edges, the sobel edge detector is applied on each sub image. The resultant edges so obtained are used to form an edge map. Morphological operations are applied on the processed edge map and further thresholding is applied to improve the performance[8].

B. Page Segmentation in OCR system For Text Extraction

Optical character recognition is an active field for recognition pattern. In this paper we tried to present how processes work in OCR system, pre-processing in OCR and document analysis. To review the process for analysis pattern from document proper page segmentation should be done. So various categories of page segmentation algorithms are mentioned which are Top Down, Bottom Up and Hybrid in this paper[9].

C. Image Segmentation Techniques

Image segmentation is the process of partitioning an image into multiple segments, so as to change the representation of an image into something that is more meaningful and easier to analyze. Several general-purpose algorithms and techniques have been developed for image segmentation. This paper describes the different segmentation techniques used in the field of ultrasound and SAR Image Processing. Firstly this paper investigates and compiles some of the technologies used for image segmentation. Then a bibliographical survey of current segmentation techniques is given in this paper and finally general tendencies in image segmentation are presented[10].

D. Existing Image Segmentation Techniques

Image segmentation is the most important part in digital image processing. Segmentation is nothing but a portion of any image and object. In image segmentation, digital image is divided into multiple set of pixels. Image segmentation is generally required to cut out region of interest (ROI) from an image. Currently there are many different algorithms available for image segmentation. Each have their own advantages and purpose. In this paper, different image segmentation algorithms with their prospects are reviewed[11].

III. PROBLEM STATEMENT

The problem in hand is such that to segment out the text content and noise from a scanned pdf of image document. To do that, we need to process the document from all the sides in such a way that we segment out all the noise at one place and all the text in another place. We can then review the segmented results. The process will include a lot of work using structuring elements in MATLAB. The main aim of the segmentation is to simplify or change the representation of the image into data or something that is more meaningful and easier to observe.

IV. OBJECTIVES

Optical Character Recognition (OCR) is the effective automated process of translating or convert an input document image (Scanned Document) into a symbolic text file (Microsoft Word Document).

The input scanned document images can come from a wide variety of media, such as newsletters, national and international journals, newspapers, magazines, memos, etc. The format or pattern of a input scanned document image can be digitally created, faxed, scanned, machine printed, or handwritten, etc.

The output symbolic text file from an OCR system can include not only the text content of the input scanned document image but also additional descriptive information, such as page layout, font size and elegance, document region kind, confidence level for the recognized characters, etc.

Page segmentation is a critical pre-processing step in an OCR system. It is the process of dividing a document image into homogeneous zones, such as zones containing only similar information i.e. text, tables, figures, or halftone images etc. In several cases, OCR system accuracy heavily depends on the accuracy of the page segmentation method.

V. RESULTS AND DISCUSSION

Following are the results for this simulation firstly the scanned image will be processed and the boundaries of all sides right, left, up and down will be detected and removed then final image detected after the noise removal process and in last the scanned document will convert in the text content that will be come in the command window of the matlab.

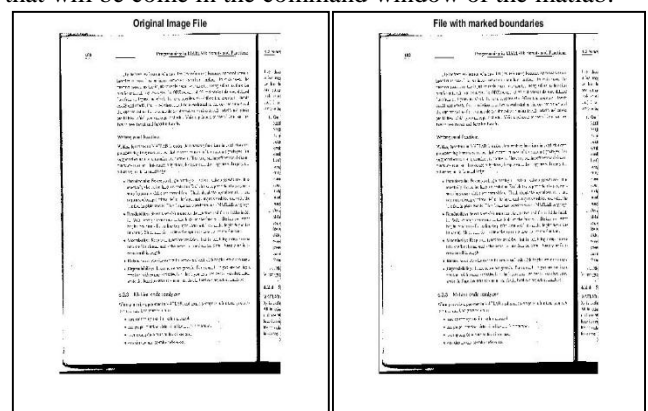


Figure 1: (a) Original scanned image (b) File with marked boundaries

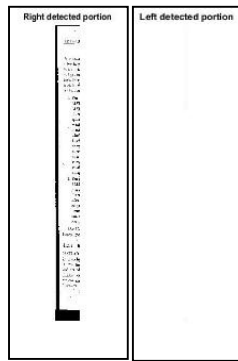


Figure 2: (a) Right detected portion (b) Left detected portion

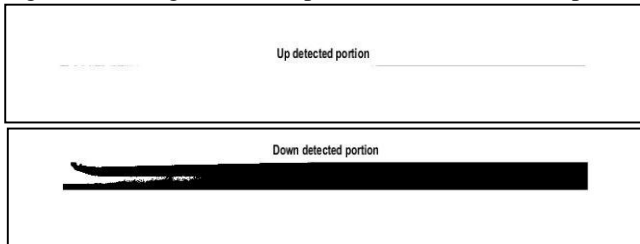


Figure 3: (a) Up detected portion (b) Down detected portion

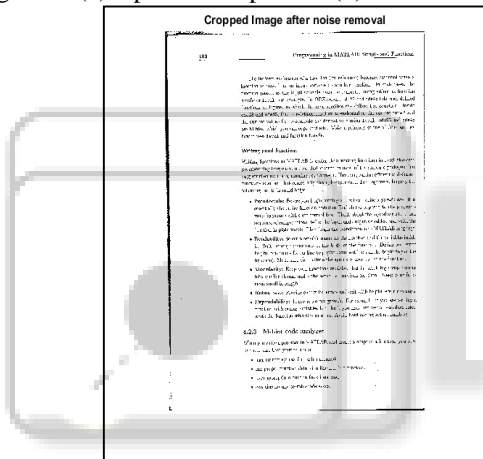


Figure 4: Cropped Image after noise removal

REFERENCES

[1] Z. Razak, K. Zulkiflee, et al., Off-line handwriting text line segmentation: a review, *International Journal of Computer Science and Network Security* 8 (7) (2008) 12–20.

[2] B. Yanikoglu, P.A. Sandon, Segmentation of off-line cursive handwriting using linear programming, *Pattern Recognition* 31 (12) (1998) 1825–1833.

[3] G. Louloudis, B. Gatos, C. Halatsis, Text line detection in unconstrained handwritten documents using a block-based Hough transform approach, in: *Proceedings of International Conference on Document Analysis and Recognition*, 2007, pp. 599–603.

[4] Wu, R. Manmatha, E.M. Riseman, TextFinder: an automatic system to detect and recognize text in images, *IEEE Trans. Pattern Anal. Mach. Intell.* 12 (1999) 1224–1229.

[5] C. Strouthopoulos, N. Papamarkos*, A.E. Atsalakis, "Text extraction in complex color documents" *Pattern Recognition* 35 (2002) 1743–1758.

[6] H. Choi and R. G. Baraniuk, "Multiscale Image Segmentation Using Wavelet-Domain Hidden Markov Models", *IEEE Transactions on Image Processing*, vol. 10(9), pp. 1309-1321, Sep. 2001

[7] Shulan Deng and Shahram Latifi, "Fast Text Segmentation Using Wavelet for Document Processing", *Proc. of the 4th WAC, ISSCI, IFMIP*, Maui, Hawaii, USA, pp. 739-744, 11-15 June 2000.

[8] Neha Gupta and V.K. Banga, "Image Segmentation for Text Extraction", *2nd International Conference on Electrical, Electronics and Civil Engineering*, pp.182-185, april 28-29, 2012.

[9] Sukhvir Kaur, P.S.Mann and Shivani Khurana, "Page Segmentation in OCR System", *International Journal of Computer Science and Information Technologies*, pp. 420-422, vol. 4 (3) , 2013.

[10] Rajeshwar Das, Priyanka and Swapna Devi, "Image Segmentation Techniques", *International Journal of Electronics & communication Technology*, pp.66-70, vol.3, 2012.

[11] Rohan Kandwal, Ashok Kumar and Sanjay Bhargava, "Existing Image Segmentation Techniques", *International Journal of Advanced Research in Computer Science & Software Engineering*, pp.153-156, vol.4, 2014