A Novel Scheme for the extraction of textual areas of a scanned document using Page Layout Segmentation Algorithm: A Review

Jyoti
M.TECH (CSE), Research Scholar
Department of Computer Science & Engineering
OM Institute of Technology and Management Hisar, India
jmehta031@gmail.com

Er. Amit Ranjan
Assistant Professor
Department of Computer Science & Engineering
OM Institute of Technology and Management Hisar, India
amitrnjn87@gmail.com

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.

I. INTRODUCTION

Scanned document image segmentation to text content and words is a difficult stage towards unconstrained handwritten document recognition[1]. As there is a drastic advancement in Computer Technology & communication technology, the modern era is entering to the information and knowledge edge. In advancement in the document system (paper, books etc), people nowadays using electronic system of documentation (PDF and word 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 done before its entry. Image segmentation theory, as digital image processing has become a very important part of people active research. Scanned image processing document image segmentation theory is an important area of research in the process it is mainly between the document image pre-processing and advanced character recognition an important link between. The comparatively effective and commonly used for document image segmentation and classification ways include threshold, and geometric data and other categories[2].

After the segmenting process, text part is detected from the scanned document and extracted for further process of conversion, earlier, text extraction techniques have been made only on monochrome documents. We can classified these techniques as bottom-up, top-down and hybrid[2]. Here, the problem of locating the textual data in an image has been addressed. Further, the extended text extraction scheme for the segmentation of document images has done. Our text extraction scheme from scanned document can identify and isolate textual regions in these kind of images. Such kind of system finds different applications in image and text database recover, automated processing and reading of the all kind of documents, and storing the same documents in digitized form[3]. Matlab is used in mathematical calculation, data analysis etc[9]. Matlab is very help to implement our ideas beyond desktop[10].

1.1 IMAGE SEGMENTATION

The process of dividing a digital image into multiple segments like (sets of pixels, also known as super pixels). The main objective of segmentation is to simplify and/or change the representation of a digital image into more meaningful and easier to observe. Image segmentation is generally used to find objects and boundaries (lines, curves, etc.) in digital images. Basically, image segmentation is the method of assigning a tag to every pixel in a digital image such that pixels with the same tag share certain characteristics of the image. In image segmentation many different algorithm approach with different prospects are available[13].

1.1.1 Need

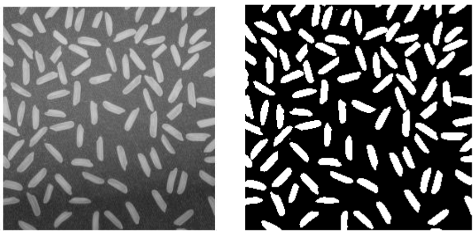
- ➢ Improving the analysis of an image when there is no direct correspondence between the image pixel properties and the type of tissue.
- ➢ Separating (labelling) the pixels of an image according to semantic content (studied structure).
- ➢ Facilitating the manipulation and visualization of the data with a computer.
- ➢ Segmentation involves the partitioning of an image or volume into distinct (usually) non overlapping regions in a meaningful way.
- ➢ It can also be thought of as a labelling operation: a label corresponding to tissue type/anatomical structure is assigned to each pixel or voxel in the image.
- ➢ It can also identifies separate objects within an image.
- ➢ Segmentation is also required for finding the regions of connected pixels with similar properties.

1.1.2 Example

Simple example: Segmentation of rice grains

Each pixel is assigned a label:

- 0 = not rice grain pixel
- 1 = rice grain pixel



Original image Segmented (binary) image

Figure1.1 Segmentation of rice grains

1.2 CURRENT IMAGE SEGMENTATION TECHNIQUE

In recent years, a lot of research is done in the field of image segmentation process. There are currently thousands of algorithm, each doing the segmentation process slightly different from another, but still there is no particular algorithm that is applicable for all types of digital image, fulfilling every objective. Thus, the technique developed for a group of digital images may not always apply to images of another class. In image segmentation if a page then update the database, if delete the page then check deletion log[11].

Currently image segmentation approach, based on two properties of an image, is divided into two categories:

- *Discontinuities based*

In this technique, subdivision of digital images are carried out on the basis of abrupt changes in the intensity of grey levels of an image. It is based on identification of isolated points, lines and edges. This include image segmentation algorithms like edge detection.

- *Similarities based*

In this category, subdivision of images are carried out on the basis of similarities in intensity or grey levels of an image. It is based on identification of similar points, lines and edges. This represents image segmentation algorithms like thresholding, region growing, region splitting and merging.

II. LITERATURE REVIEW

This section describes the literature review about various papers from various journals studied. It presents the review of earlier work done. Literature review discusses the published information in a particular subject area within a certain time period. It can be just a simple summary of sources, but it usually has an organization patterns and combines both summary and synthesis. A summary is a recap of the information of the source but synthesis is a reorganization of that information. It seeks to describe, summarize, evaluate, clarify and integrate the content of primary reports in graphical authentication. In which firstly investigate the paper and compiles then finally general tendencies in image segmentation are presented[12].

2.1 Handwritten Document Image Segmentation Into Text Lines And Words

Two main methods to extract text lines and words from handwritten document are presented in this paper. The segmentation technique or algorithm is based on locating the optimal succession of text content and gap areas within vertical zones by applying Viterbi algorithm. Then, a text-line separator drawing technique is applied and in last the connected components are assigned to text lines. Word segmentation is based on a gap metric that exploits the objective function of a soft-margin linear SVM that separates successive connected components. The algorithms tested on the bench marking datasets of ICDAR07 handwriting segmentation contest and outperformed the participating algorithms[1].

2.2 Enhanced Techniques For Pdf Image Segmentation And Text Extraction

Convert text content from the PDF images is a challenging problem. The text data present in the PDF images contain certain useful information for automatic annotation, indexing etc. However variations of the text due to differences in text style, font, size, orientation, alignment as well as complex structure make the problem of automatic text extraction extremely difficult and challenging job. This paper presents two techniques under block-based classification. After a brief introduction of the classification methods, two methods were enhanced and results were evaluated. The performance metrics for segmentation and time consumption are tested for both the models[2].

2.3 Text Extraction And Document Image Segmentation Using Matched Wavelets And MRF Model

In this paper, we have proposed a novel scheme for the extraction of textual areas of an image using globally matched wavelet filters. A clustering-based technique has been devised for estimating globally matched wavelet filters using a collection of ground truth images. We have extended our text extraction scheme for the segmentation of document images into text, background, and picture components (which include graphics and continuous tone images). Multiple, two-class Fisher classifiers have been used for this purpose. We also exploit contextual information by using a Markov random field formulation-based pixel labeling scheme for refinement of the segmentation results. Experimental results have established effectiveness of our approach[3].

2.4 Segmentation of Text from Image Document

Segmentation of text from image documents has many important applications such as document retrieving, object identification, detection of vehicle license plate, etc. It is very popular research field in recent years. In this paper, we employ Symlet wavelet and 2-mean classification for segmentation of text from image document. We have used morphology operation like as dilation and erosion in post processing. Proposed method for text segmentation from image document has been implemented in MATLAB[4].

2.5 Text Detection From Documented Image Using Image Segmentation

The Segmentation subdivides an image into its constituent region or objects. The level to which the subdivision is carried depends on the problem being solved. That is segmentation should stop when the object of interest in an application have been isolated. The segmentation of nontrivial Images is one of the most difficult tasks in image processing. Segmentation accuracy determines the eventual success or failure of computerized 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 problems encountered in the segmentation, these includes the difference in the skew angle between lines, characters or even along the same text line, adjacent text line, overlapping words and touching characters[5].

2.6 Fast Document Segmentation Using Contour And X-Y Cut Technique

This paper describes fast and efficient method for page segmentation of document containing non rectangular block. The segmentation is based on edge following algorithm using small window of 16 by 32 pixels. This segmentation is very fast since only border pixels of paragraph are used without scanning the whole page. Still, the segmentation may contain error if the space between them is smaller than the window used in edge following. Consequently, this paper reduce this error by first identify the missed segmentation point using direction information in edge following then, using X-Y cut at the missed segmentation point to separate the connected columns. The advantage of the proposed method is the fast identification of missed segmentation point. This methodology is faster with fewer overheads than other algorithms that need to access much more pixel of a document[6]. In which discrete wave transform is used for image preprocessing [7].

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.

3.1 OBJECTIVES

1. 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).In page segmentation Top Down, Bottom Up approaches for optical character recognition[8].

2. The input scanned document images can come from a wide variety of media, such as newslatters, 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 style, document region type, confidence level for the recognized characters, etc.

IV. REFERENCES

[1] VassiliPapavassiliou, ThemosStafylakis, VassilisKatsourosaandGeorgeCarayannis,“*Handwritten document image segmentation into text lines and words*”,National Technical University of Athens, School of Electrical and Computer Engineers, pp.369-377,2010.

[2] D.Sasirekha and Dr.E.Chandra,“*Enhanced Techniques for PDF Image Segmentation and Text Extraction”,* International Journal of Computer Science and Information Security,vol.10, no. 9,september

[3] Sunil Kumar, Rajat Gupta and Nitin Khanna,“ *Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model*”, IEEE Transactions on Image Processing,pp.2117-2128 vol. 16, no. 8, august 2007.

[4] Ankush Gautam,“ *Segmentation of Text From Image Document*”,International Journal of Computer Science and Information Technologies, pp. 538-540,vol. 4 (3), 2013.

[5] Santosh and Dr. Jenila Livingston L.M,“*Text Detection From Documented Image Using Image Segmentation*”,International Journal of Technology Enhancements and Emerging Engineering Research, pp.144-148,vol.1,2013.

[6] Boontee Kruatrachue, Narongchai Moongfangklang and Kritawan Siriboon,“*Fast Document Segmentation Using Contour and X-Y Cut Technique “,*World Academy of Science, Engineering and Technology,pp.27-29,2007.

[7] 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.

[8] 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.

[9] cimss.ssec.wisc.edu/wxwise/class/aos340/spr00/whatismatlab.htm

[10] www.mathworks.com/help/matlab/

[11] http://en.wikipedia.org/wiki/image_segmentation.

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

[13] 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.

V. AUTHOR’S BIOGRAPHIES

Jyoti is a MTECH student in department of Computer Science & Engineering from OM Institute of Technology and Management, Hisar (Haryana). She received B.TECH degree in Computer Science & Engineering from HCTM Technical Campus, Kaithal (Haryana). Her research include extraction of textual areas of scanned documents using page layout segmentation algorithms.