Empirical Performance Evaluation Methodology and its Application to Page Segmentation Algorithms: A Review

Pinky Gather[1], Avininder Singh[2]

[1] M.Tech Student, [2]Assistant Professor
[1,2] *Department of Computer Science and Applications,*
Ch. Devi Lal University, Sirsa(Hry), India.

***Abstract*—Text extraction plays an important role in computer vision for providing useful and valuable information. The main component of document image analysis is text line page segmentaion.There are many factors on which document image depends i.e. background, color, sizes, orientation and touching text lines. The area of background and photo is taken as an approach for scanned document. In printing industry such classification algorithms can be used for object oriented rendering and enhanced scanning. We propose a page-layout-segmentation technique to extract text from scanned documents.**

***Index Terms*—page segmentation, X-Y Cut Technique, document image.**

I. INTRODUCTION

Document image segmentation to text lines and words is a critical stage towards unconstrained handwritten document recognition[1]. With the drastic advancement in Computer Technology & communication technology, the modern society is entering to the information 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[2].

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 an important part of people active research. Image processing document image segmentation theory is an important research topic in the process it is mainly between the document image pre-processing and advanced character recognition an important link between. The relatively effective and commonly used for document image segmentation and classification methods include threshold, and geometric analysis and other categories[2].

After segmenting, Text part is detected and extracted for further process, earlier, text extraction techniques have been developed only on monochrome documents. These techniques can be classified as bottom-up, top-down and hybrid[2]. Here, we address the problem of locating the textual data in an image. Further, we have extended text extraction scheme for the segmentation of document images. Our text extraction scheme can identify and isolate textual regions in these kind of images. Such a system finds applications in image and text database retrieval, automated processing and reading of documents, and storing the documents in digitized form[3].

Segmentation accuracy determines the eventual success or failure of computerized analysis procedures[5]. 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].

Page segmentation is the process to identify the areas of interest in the image of a document page. For a conventional document page with material printed in dark ink on a light colored paper, the areas of interest in the (binary) image will be neighbourhoods of black pixels. Page segmentation produces a description of the geometrical aspects of the areas of interest. The most common aspects are spatial extent and position on the page. Page segmentation can be thought of as a mapping from the pixel-based image data to a description of the areas of interest[6].

II. LITERATURE REVIEW

A. Handwritten document image segmentation into text lines and words

Two novel approaches to extract text lines and words from handwritten document are presented. The line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones by applying Viterbi algorithm. Then, a text-line separator drawing technique is applied and finally 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 out performed the participating algorithms[1].

B. Enhanced Techniques for PDF Image Segmentation and Text Extraction

Extracting text objects from the PDF images is a challenging problem. The text data in the PDF images contain certain useful in formation 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].

C. 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].

D. 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, weemploy 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].

E. 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].

F. 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]

G. 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 gray scale 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[7].

III. PROBLEM STATEMENT

The problem in hand is such that we need to segment out the text and noise from a scanned 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 review the segmented results. The process will include a lot of work using structuring elements in MATLAB. The aim of the page segmentation is to convert the image into more meaningful representation that is easier to analyze.

IV. TOOLS AND TECHNOLOGY USED

A. The MATLAB System

The MATLAB Language is a high-level matrix/array language with control flow statements, functions, data structures, input/output, and object-oriented programming features. It allows both “programming in the small” too rapidly create quick and dirty throw-away programs, and “programming in the large” to create complete large and complex application programs. The language features are organized into six directories in the MATLAB Toolbox[9].

MATLAB is a high-level language and interactive environment for numerical computation, visualization, and programming. Using MATLAB, we can analyze data, develop algorithms, and create models and applications. The language, tools, and built-in math functions enable you to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages, such as C/C++ or Java. We can use MATLAB for a range of applications, including signal processing and communications, image and video processing, control systems, test and measurement, computational finance, and computational biology[10].

V. OBJECTIVES

1. Optical Character Recognition (OCR) is the automated process of translating an input document image into a symbolic text file.
2. The input document images can come from a large variety of media, such as journals, newspapers, magazines, memos, etc. The format of a 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 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.
3. Page segmentation is a crucial preprocessing 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 many cases, OCR system accuracy heavily depends on the accuracy of the page segmentation algorithm.

VI. CONCLUSION

We are locating the text part based on the textural attributes by presenting the novel technique using GMWs. The filtering and the feature extraction operations account for most of the required computations however, our method is very simple, computationally less expensive and efficient. Compared to other existing methods the dimensionality, and, so, the computation of the feature space, is considerably reduced. We have applied our algorithm on several structured and highly unstructured images with complex backgrounds and obtained encouraging results.

REFERENCES

[1] VassiliPapavassiliou, ThemosStafylakis, VassilisKatsourosa and GeorgeCarayannis, "*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 2012.

[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/