ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

Performance Analysis of Time in Line and Word Segmentation using Different Fonts

Shashi Kant[1], Mrs. Sini Shibu[2]

M.Tech Research Scholar, Department of CSE, NRI Institute of Information Sc. & Tech., Bhopal[1]

Asst. Professor, Department of CSE, NRI Institute of Information Sc. & Tech., Bhopal[2]

Abstract: Segmentation is a widely used process in current-day image processing. Various segmentation processes exist to segment lines and words. Header lines are detected and converted as straight lines. Text line segmentation is avn important step because inaccurately segmented text lines will cause errors in the recognition stage. Text characteristics can vary in font, size, and orientation. This paper introduces a comparative analysis of time of the segmentation process that is carried out on various style text to find the time taken in the segmentation process. The text is segmented into lines, lines into words. Here we make a comparative study of time consumed during line and word segmentation using different style fonts and find the time in which fonts type, both segmentation take minimum amount of CPU time. Here we proposed a method that supports all segmentation process(line and word segmentation) to retrieve text, make boundary boxes to the fully lines, words and perform segmentation. It consists of recording the start time of the process as well as end time of the process and find the time difference. This process is applied to different fonts of text to make a comparative analysis.

Keywords: Image processing, Printed/Handwritten document analysis, Text font time analysis, segmentation, line segmentation and word segmentation.

I. INTRODUCTION

Segmentation of a printed text image into its basic entities, namely, text lines words and characters, is considered as a non-trivial problem to solve in the field of printed document recognition and processing. The difficulties that arise in printed documents make the segmentation procedure a challenging task. Different types of difficulties are encountered in the text line segmentation and word segmentation. In case of text line segmentation procedure, major difficulties include the difference in the skew angle between lines on the page or even along the same text line, overlapping words and adjacent text lines touching. Printed text recognition is difficult task compared to machine printed different fonts style line and word recognition in the area of text recognition.

A lot of research is done on the printed English text, but less work has been done on the printed English text line and word recognition. It has four major steps such as pre-processing, segmentation recognition and CPU time consuming. Segmentation is the important step. Segmentation also contains three major steps such as line segmentation, word segmentation and character segmentation. If we fail in doing line and word segmentation then entire segmentation process goes wrong. No more research has been done in the past on line and word segmentation of machine printed different fonts style texts. This approach basically limits the processing time accuracy achieved by the text recognition system.

Text recognition, usually shortened to TR, is the electronic translation of scanned of typewritten or printed text images into machine-encoded/computer-readable text.

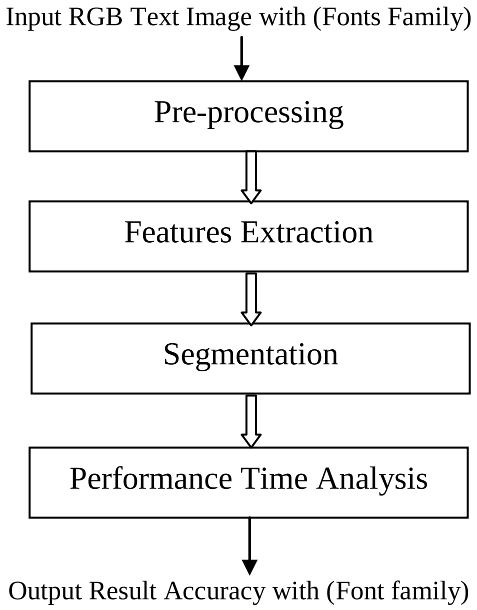
Input RGB Text Image with (Fonts Family)
Pre-processing
Features Extraction
Segmentation
Performance Time Analysis
Output Result Accuracy with (Font family)

Fig. 1 Text Information Extraction System

The popularity of TR has been increasing each year with the advent of fast microprocessors providing the base for vastly improved recognition techniques. Now, there have been tremendous improvements in increasing both effective read rates and accuracy of line and word recognition. Desktop TR scanners can recite typewritten data into a computer at rates up to 2400 words per minute! As we all know it is used for many types of data entry whether bank statement, business card, passport documents, invoices, receipts, mail, or any number of printed records.

DOI 10.17148/IJARCCE.2015.4690

II. SEGMENTATION BASIC

Segmentation is one of the most important phases in Text recognition process. Segmentation is the process of segmenting the whole input document image into recognizable units. Segmentation can be useful for vision clarity, medical operations, because of due to this every large image can be segmented manually, and segmented small region helps to judge every pixel of image clearly for the further more activities.

The segmentation process is divided into three types.
A. Line segmentation
B. Word segmentation
C. Character segmentation

A. Line Segmentation

First step of segmentation process is segmenting the text region into lines, also called as line segmentation. First we need to calculate header lines and base lines for the Line segmentation Header lines are rows with maximum number of black pixels and base lines are rows with minimum number of black pixels. Finding header line is a challenge because of skew in headline.

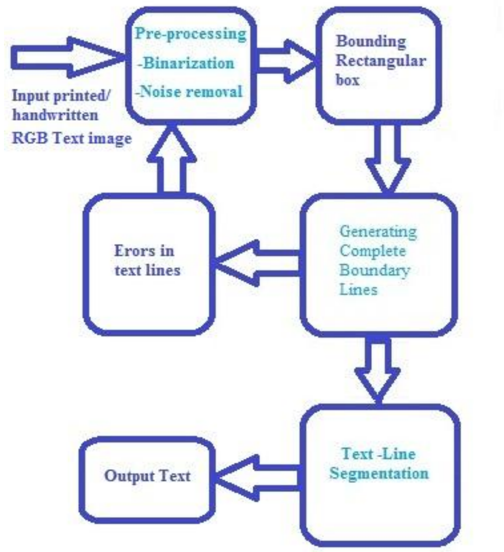

Fig. 2 Flowchart for line segmentation process

A text line is separated from the input text image by applying horizontal projection method for compute horizontal text lines, we can compute corresponding horizontal histogram of an image in which each and every row of input text image can be separate individually. Since in English script, there always present a gap between the every character they are not properly attached with each other.

B. Word Segmentation

Word segmentation is easier than line segmentation and character segmentation. Space between two words is generally more than three pixels. Words are segmented by the projection based method. Word segmentation is the problem of dividing a string of written language into its component words. In English and many other languages using some form of the Latin alphabet, the space is a good approximation of a word divider (word delimiter). However the equivalent to this character is not found in all written scripts, and without it word segmentation is a difficult problem. Languages which do not have a trivial word segmentation process include Chinese, Japanese, where sentences but not words are delimited,

Proposed Algorithm-
Step 1: Read the RGB input text image.
Step 2: Detecting input noise and Removing noise using noise filter.
Step 3: Feature Extracting , considering binarized text lines. and perform some morphological and projection profile method.
Step 4: Creating boundary box region to grouping the individual words from the filtered text line.
Step 5: Perform word segmentation for segmented result as well as processing CPU time analysis for different fonts style.
Step 6: If (word segmentation = true && Font based CPU time analysis = true). Found the expected output
else error not done, Repeat step 3 to 5 and exit.

However we are performing the line and word segmentation on different fonts like Times New Roman, Arial Rounded, Calibri , verdana text, Cambria and Berlin sans text etc. by using sample text images.

C. Character Segmentation

Character segmentation is to separate individual characters from text word. In printed or scanned text header line is not straight line. Character segmentation is a process of grouping and segmenting every single unit of text and symbols from the input image or document.

The rest of this paper is organized as follows. Section III presents some of the related works in text localization, and segmentation. Section IV describes the implementation of the two segmentation approaches. Experiments and result analysis are discussed in Section V. Section VI concludes the paper with future recommendations.

III. LITERATURE REVIEW

Text detection is usually performed on text features such as color, edge, texture. Text detection is classified into three main categories: Color based, Edge based and Texture based. G. Seni and E. Cohen [2] used External Word Segmentation of Off-Line Handwritten Text Lines. Marcin Pazio et al. Nallapareddy Priyanka [3] provide a new concept for Line and Word Segmentation Approach for Printed Documents, this includes printed or scan English text image binaraized to make segmentation. [6] used color based approach for text detection. Segmentation is followed by connected component analysis. Azadboni [12] proposed a two stage algorithm for text localisation. In first stage, image is preprocessed to remove noise and increase the contrast, followed by projection profile analysis to extract text blocks. In the second stage, extracted text blocks are verified using SVM classifier. In [4] U. Mahadevan and R.C. Nagabhushanam, prepare a Gap Metrics for Word Separation in Handwritten Lines, to use it for extraction words. Finally segmentation is performed to extract character pixels. W. Huang et al. [13] used Stroke Feature Transform (SFT) followed by a text component classifier and a text-line classifier, sequentially to extract text regions. Finally, text regions are located by the text-line confident map. In [7] [8] texture based approach is discussed in which text is considered as a specific form of texture. In [9] F. Hones and J. Litcher, provide a machine based method for Layout extraction of mixed mode documents. In [10][11], edge based approach for text detection is discussed in the literature. Zramdini et al. [1] introduced font style detection method by calculating the CPU time taken in the horizontal profile of the whole text block. N. Sharma et al. [5] also presented a technique for improving the recognition accuracy of Hindi OCR system by developing the concept for detection of bold, italic word of different fonts. It is evident from the literature survey that none of the researchers had focussed on detecting total time is to be taken in line and word segmentation using different fonts based and all capital Fonts pattern words in printed in Arial Rounded, Times New Roman, etc script till date. This has motivated us to propose a two-stage font invariant detection technique for detecting all the above mention type font time accuracy during processing

IV. COLOR BASED AND FONTS BASED TEXT SEGMENTATION

This section discusses segmentation approach for text detection. In text, letters should have uniform color and fonts family within a text string. This property is used in color based text detection and fonts based time analysis approaches. The general algorithm for color based and font based line and word text detection approach is as follows. Initially Text image is segmented.

Step 1: As simply color image is processed and make transforming it into rgb2gray for extracting the color.

Step 2: Binarization the processed image and noise removing and applying horizontal projection method for read text in rows.

Step 3: Grouping rows and perform on input text image to find the line and word segments.

Step 4: Calculating CPU time, taken by the overall segmentation process on different fonts families.

Here we discuss some texts image machine printed or scan RBG text image with different types of fonts namely- arial,times new roman, calibri etc in Fig. 3. As we know that for the time performance and analysis during segmentation to find the overall fonts based segmentation process performed in less time to optimize the result. For this process a routine process is to be as follows.

A. Pre-processing.
B. Feature Detection
C. Result Analysis.

A. Pre-processing

In pre processing the printed or scan text image is converted into ready to use format or the noise in the image is reduced by using these methods, in the scanned image the noise would be like word shapes are not accurately scanned document was not associated properly, the document is tilted to rare degrees, edges are not smooth, multiple colored characters, So, the remove such noise, we have to apply some logics as computer doesn't know the difference between the noise and the accurate character. To remove the noise the following algorithms can be applied to make it ready to use for line and word recognition.

De-skew

This method is used for proper alignment of document which was not aligned properly after scanning, it may provide few degrees clockwise or anti-clockwise in order to make lines of text perfectly horizontal or vertical.

Binarization

This technique is used to convert an image from RGB color to grayscale to black-and-white to make it appropriate for extracting line and word recognition and it is also known as "binary image" because there are only two colors left in the image after the binarization i.e. "BLACK" and "WHITE". The pixel consists value 1 for black region and value 0 for white region.

B. Feature Detection

After the Binarization of the 2D text image in to binary pixels values (0, 1). The header able to read on region and rectangular box is to formed in to each of line after that similar for word. And this is also known as intelligent word recognition (IWR) though this is much more sophisticated way of spotting characters. Most omni font OCR programs i.e. ones that can recognize printed text in any font work by feature detection rather than pattern recognition.

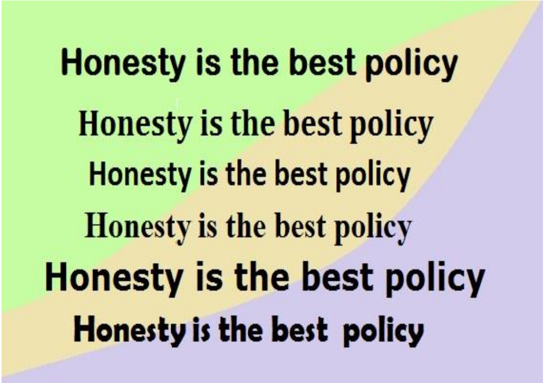
Honesty is the best policy
Honesty is the best policy
Honesty is the best policy
Honesty is the best policy
Honesty is the best policy
Honesty is the best policy

Fig. 3. Illustration of six fonts family text namely- Arial Rounded, Cambria, Calibri, Verdana, Times New Roman and Berlin sans.

V. RESULT AND DISCUSSION

The performance analysis of time during different fonts text segmentation approaches in line and word detection and segmentation one by one, is carried out in a set of 12 text images selected from printed or scanned documents. The selected images contain both images with uniform color intensity and non uniform intensity. The algorithms are implemented in MATLAB R2010a. A subjective evaluation is performed on the segmentation results. Table 1.shows the output of line segmentation and Table 2. Shows the output of word segmentation from different font type text where all containing image are processed and find 100 lines and more than 500 words to be segmented correctly and arise some errors during process and produced resultant output to find the accuracy of 88.3 % in line segmentation and similar 90.4 % in word segmentation, efficiently the Times new Roman based segmentation. In RGB based segmentation, changes in intensity causes part of text (line and word) to merge into background. Time analysis of this existence proposed method is also by a graphical form in Fig. 4. below.

Font Family	Line segmentation performance analysis			
	Text lines	Correctly Segmented	Error	Time (CPU) Accuracy
Arial Rounded	100	98.2 %	1.8 %	87.8 %
Times New Roman	100	98.5 %	1.5 %	88.3 %
Cambria	100	97.5 %	2.5 %	86.9 %
Calibri	100	96.4 %	3.6 %	85.6 %
Verdana	100	97.8 %	2.2 %	86.5 %
Berlin Sans	100	95.3 %	4.7 %	84.7 %

Table 1. Output result of line segmentation.

Font Family	Line segmentation performance analysis			
	Text words	Correctly Segmented	Error	Time (CPU) Accuracy
Arial Rounded	500	97 %	3 %	90 %
Times New Roman	500	97.5 %	2.5 %	90.4 %
Cambria	500	95 %	5 %	87.7 %
Calibri	500	95.5 %	4.5 %	86.1 %
Verdana	500	96 %	4 %	88.2 %
Berlin Sans	500	94 %	6 %	85.3 %

Table 2. Output result of word segmentation

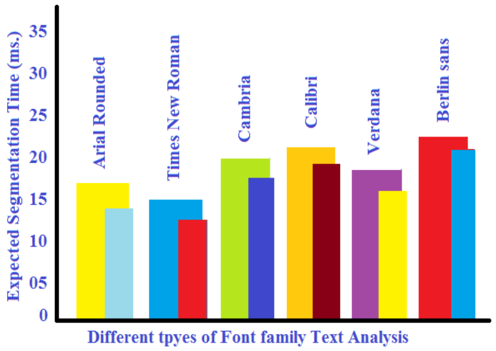

Fig. 4. Analysis of time accuracy to each font

VI. CONCLUSION

From the above both tables and accuracy graph, we can say that Times New Roman and Arial Rounded fonts are better than other to understand and detect the text for line and word segmentation. It takes less time for segmentation and make effective segmentation on text-lines and words. As compared to all the different fonts based text recognition (FTR) techniques for segmentation (line and word), we found that every font consist of different variance and orientation. We implement the code in MATLAB and perform font based text segmentation on both line and word. Thus, in the proposed method we found that segmentation approach can provide best result and analysis with Times New Roman and Ariel Rounded fonts..

REFFERENCES

[1] A. ZramdiniS, R. Ingold, “Optical font recognition using typographical features,” In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 877-882, 1998.

[2] G. Seni and E. Cohen, ªExternal Word Segmentation of Off-Line Handwritten Text Lines,”Pattern Recognition,vol. 27, no. 1, pp. 41-52, 1994.

[3] Nallapareddy Priyanka, Srikanta Pal, Ranju Mandal “Line and Word Segmentation Approach for Printed Documents”, IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” RTIPPR, pp.31-33, 2010.

[4] U. Mahadevan and R.C. Nagabhushanam, ªGap Metrics for Word Separation in Handwritten Lines,ºProc. Third Int'l Conf. Document Analysis and Recognition,pp. 124-127, Montreal, (ICDAR '95), Aug.1995.

[5] N. Sharma, M. Khandelwal, “Detection of Bold Italic and Underline Fonts for Hindi OCR”, In: International Journal of Computer Trends and Technology (IJCTT), vol. 4, Issue 8, pp. 2425-2428, 2013.

[6] Marcin Pazio, Maciej Nied´zwiecki, Ryszard Kowalik, Jacek Lebied´z, “Text Detection System for the Blind”, 15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 3-7, 2007,pp.272-276.

[7] S. A. Angadi & M. M. Kodabagi, “Text Region Extraction from Low Resolution Natural Scene Images using Texture Features”, International Journal of Image Processing, vol. 3, issue 5, pp. 229-245, 2009.

[8] Kwang In Kim, Keechul Jung, And Jin Hyung Kim,“Texture-Based Approach For Text Detection In Images Using Support Vector Machines And Continuously Adaptive Mean Shift Algorithm” , IEEE Transactions On Pattern Analysis And Machine Intelligence , Vol. 25, No. 12, 2003.

[9] F. Hones and J. Litcher, “Layout extraction of mixed mode documents”, Machine Vision Application, vol. 7, pp. 237–246, 1994.

[10] Andrej Ikica, Peter Peer, “An improved edge profile based method for text detection in images of natural scenes”, EUROCON- International Conference on Computer as a tool (EUROCON) 2011 IEEE: 1-4.

[11] Jing Zhang and Rangachar Kasturi, “Text Detection Using Edge Gradient and Graph Spectrum”,2010 International Conference on Pattern Recognition, 2010 IEEE ,pp 3979-3982.

[12] Azadboni, M.K. ; Behrad, A.“Text detection and character extraction in color images using FFT domain filtering and SVM classification”, 2012 Sixth International Symposium on Telecommunications (IST), 2012 IEEE ,pp 794-799. DOI: 10.1109/ISTEL.2012.6483094

[13] Weilin Huang, Zhe Lin , Jianchao Yang ,Jue Wang, “Text Localization in Natural Images Using Stroke Feature Transform and Text Covariance Descriptors”,International Conference on Computer vision (ICCV), 2013 IEEE ,pp 1241-1248.

[14] R. C. Gonzalez, R. E. Woods, “Digital Image Processing”, vol. I. Prentice-Hall, India (1992).

DOI 10.17148/IJARCCE.2015.4690