IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 1, Ver. I (Jan - Feb. 2015), PP 15-21
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

An Efficient Segmentation Technique for Machine Printed Devanagiri Script: Both Line &Word Segmentation

[1]Siba Kumar Panda, [2]Smruti Snigdha Pani, [3]Biranchi Narayan Panda
[1]*Assistant Professor, Dept. of ECE Centurion University of Technology and Management, Jatni, Odisha*
[2]*M.Tech Scholar, Dept. of ETC Veer Surendra Sai University Of Technology,Burla,Odisha*
[3]*B.Tech Scholar, Dept. of EEE, RIT, Berhampur*

***Abstract*:** *Segmentation technique plays a major role in scripting the documents for extraction of various features. Many researchers are doing various research works in this field to make the segmenting process simple as well as efficient. In this paper a simple segmentation technique for both the line and word segmentation of a script document has been proposed. The main objective of this technique is to recognize the spaces that separate two text lines.For the Word segmentation technique also similar procedure is followed. In this work ,three different scanned document have been taken as input images for both line and word segmentation techniques. The results found were outstanding with average accuracy for both line and word. It provides 100% accuracy for line segmentation and 100% for line segmentation as well. Evaluation results show that our method outperforms several competing methods.*

Key Words: *Segmentation, Line Segmentation, Word Segmentation, Devanagiri Script, Machine learning, shirorekha, white rows, white columns, consecutive white rows (CWR) threshold*

I. Introduction

The technique word segmentation is an imperative problem in many natural language processing tasks. In the paradigm of speech recognition there is no precise word boundary information given within a continuous speech statement or in interpreting written languages such as Chinese, Japanese and Thai where words are not enclosed by white-space but instead must be contingent from the basic character sequence. There is a small difference between the terms word breaking and word segmentation. Word breaking is defined as the process of segmenting known words that are predefined in a lexicon. Similarly Word segmentation is defined as the process of both glossary word segmentation and unknown word or new word revealing. For unsupervised learning in morphological analysis automatic word segmentation plays a vital role. Mounting a morphological analyzer by hand can be costly and time consuming for a new language and also it requires a great effort by highly specialized experts.

Consequence of segmentation technique accumulating sporadically at its practical application base is escalating rapidly.For numerous processes it is a primary stage such as machine recognition of language script. Segmentation is also used to extract various useful features of a document. The most important application of segmentation is detection of the script that is printed in a document, to recognize it as English or French etc. This paper mainly concentrates upon the segmentation of Devnagiri script, which is an Indian script for writing number of languages like Hindi, Marathi, Sanskrit, Sindhi and Nepali languages.

II. Related Works

In order to develop very simpler and efficient techniques for both line and word segmentation continuous researches are going on in the field of image processing and signal processing. A few important techniques for the segmentation techniques includes thinning approach based segmentation[5], segmentation using histogram method[1], header line and base line detection technique [2] , Hough transformation based technique [6], technique of smearing [7], grouping method [8], graphical approach [9], CTM (Cut text Minimum) method [10], Block covering method [11] ,text line identification approach[12]. Each method has its own importance .This section describes few important methods proposed by many researchers.

Document segmentation [1] is one of the serious phases in machine recognition of any language.Correct segmentation of entity symbols decides the accuracy of character recognition technique.It is help in decomposition of image of a sequence of characters into sub images of individual symbols by segmenting both lines and words. Devnagari is the most popular script in India. It is used for writing Hindi, Marathi, Sanskrit and Nepali languages.Moreover, Hindi is the third most popular language in the world. Devnagari documents consist of vowels, consonants and various modifiers. Hence proper segmentation of Devnagari word is a challenging task. A simple histogram based approach to segment Devnagari documents is proposed in this paper.

This paper provides the description about a new segmentation technique based on structure approach for Handwritten Hindi text [2]. Segmentation is one of the major stages of character recognition. The handwritten text is separated into lines, lines into words and words into characters. The errors in segmentation propagate to recognition. The performance is evaluated on handwritten data of 1380 words of 200 lines written by 15 different writers. The overall results of segmentation are very promising.

There are about 300 million people in India who speak Hindi and write Devnagari script. Research in Optical Character Recognition[4] is very popular for its application potential in banks, post offices, defense organizations and library automation etc. However most of the OCR systems are available for European texts.This paper describes a technique based on OCR System for different five fonts and sizes of printed Devnagari script using Artificial Neural Network. The recognition rate of the proposed OCR system with the image document of Devnagari Script has been found to be quite high.

An optical character recognition (OCR) system may provide a solution to the data entry problems[9], a diadvantage for the data processing industry. Therefore, OCR systems are being developed for almost all major languages and Arabic language is no exception to it. During the past three decades, significant research and development works have been done towards the development of an efficient Arabic optical character recognition (ACR) system.This paper mainly focuses on the comprehensive review of ACR techniques and evaluate the status of the ACR system development and an up to date bibliography.

(a).Histogram Method

This method mainly focuses on the calculation of pixel histogram.In this method a Y histogram projection is performed which results in text line position. In order to divide a line into different regions a threshold is applied. After applying a threshold another threshold is used to eliminate the false lines. However, this procedure causes some loss on the text line area. Hence, for minimizing the effect recovery method is proposed.

(b).Header line and base line detection Technique

This technique calculates the header line and base line of a text document for line segmentation. Header lines are nothing but the rows with maximum number of black pixel and base lines are the rows with minimum number of black pixel.Till now researchers are detecting the header line by finding the rows with maximum pixel density, but it cannot be applied suitably for the skew variable test.

(c).Hough transformation Technique

This technique help in explaining the procedure for the detection of collinear edge pixel even if they are isolated. This method plays an important role in finding lines in noisy images where local information around each edge pixel is unreliable or unavailable. The disadvantages lies in this method is that it requires a relatively large amount of memory and a long computation time.

(d).Technique of Smearing

The best example of Smearing technique is Run length smoothing algorithm. According to this technique the consecutive black pixels along the horizontal directions are smeared that is the white space between them is filled with black pixel if their distance is within a predefined threshold. Text line patterns are found by building a fuzzy run-length matrix, at each pixel, the fuzzy run-length is a maximal extent of the background along the horizontal direction.

(e).Grouping approach

Elementary line segments are obtained by linking edge pixel and approximating them to piecewise straight line segment. These ELS are used as input to this approach. Adjacent line segments are grouped according to some grouping criteria and replaced by new line segment. This process is repeated until no new line segment occurs. However, this approach does not work when most of the edge pixels are isolated or when the ELS's are perturbed severely by noises, rendering the data almost useless and this process is purely local. Repetition of locally optimal grouping of line segment does not guarantee their globally optimal grouping.

(f).Gradient based approach

In this approach an input image is assumed to be a gray scale image. Gradient magnitude and orientation of each pixel are explicitly used to group the pixels. The performance comparison is not easy in this approach.

This paper is organized as follows: Section I demonstrate about the introduction part. Section II explains about the related works. Section III describes briefly about the Specialty of Devnagiri script. Section IV

describes the proposed work for both line and word segmentation technique. Section V narrates the comparison analysis & result discussion and finally conclusions are discussed in Section VI.

III. Speciality Of Devnagiri Script

The presence of eighteen official languages in india makes the country as a multi-lingual country due to the presence of a typical letter for each of the phonemes in Indian languages. The national language of India obeyed as Hindi is written in the Devnagari script.Moreover Devnagari script is also used for writing various languages like Marathi, Sanskrit and Nepali. We all knows that Hindi is the third most popular language in the world [1] which is spoken by more than 500 million people in the world.The basic character in Devnagari script has 11 vowels and 33 consonants are as shown in the figure1.The Vowels can be written as independent letters or by using a variety of circumflex marks.When vowels are written in this way they are known as modifiers and the characters so formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shaped clusters are known as compound characters. These types of basic characters, compound characters and modifiers are present not only in Devnagari but also in other scripts. All the characters have a horizontal line at the upper part, known as Shirorekha. In continuous handwriting, from left to right direction, the shirorekha of one character joins with the shirorekha of the previous or next character of the same word. In this fashion, multiple characters and modified shapes in a word appear as a single connected component joined through the common shirorekha. Also in Devnagari there are vowels, consonants, vowel modifiers and compound characters, numerals. Moreover, there are many similar shaped characters

An ancient Indian script called Devnagiri which is widespreadly used to write some of the popular Indian languages.The national language Hindi is used as oral language by more than 500million people[3].Hindi is written from left to right.The presence of Shirorekha makes the Devnagiri script an unique standard in the segmentation process. Shirorekha is drawn over all the characters of a concerned word. A Devnagiri text line can be classified into three different zones with reference to position of the header line.The zones which resides over the header line is known as upper zone and the zone which hosts basic characters is known as middle zone. Similarly the zone which contains vowels or modifiers is known as lower zone.

Figure.1

Figure.2

IV. Proposed Scheme

The proposed scheme for space recognition approach for line and word segmentation fundamentally works with the recognition of the position of the spaces present in a given script document. The design flow for line segmentation and word segmentation is as shown in figure.3 and figure .4 respectively. This method works as follows:

Line Segmentation:

The mechanism of line segmentation explains in finding the white pixels in each row of a given text document taken in jpg format. In order to start with the line segmentation process, it starts with the rows which has its all the elements as white pixels that is the row which does not have any of its element with pixel value zero.This kind of row as a white rows.A simplified design flow of line segmentation is as shown in figure.3.Then appearances of the successive white rows (CWRs) through out the image are taken into consideration.A threshold value is taken number of times for which the white rows appear serially.This threshold value called as CWR threshold determines the spaces present in between text lines of the concerned image document.Particular reflection has to be paid towards upper zone and lower zone of the Devnagiri text line since some spaces separates these zones from the middle zone. hence they are not derelict by the processor as white spaces. Now without any nuisance we are capable of doing the segmentation part.The most effectiveness of the assumed threshold number is that it will guide the MATLAB processor for segmenting the exact text parts .

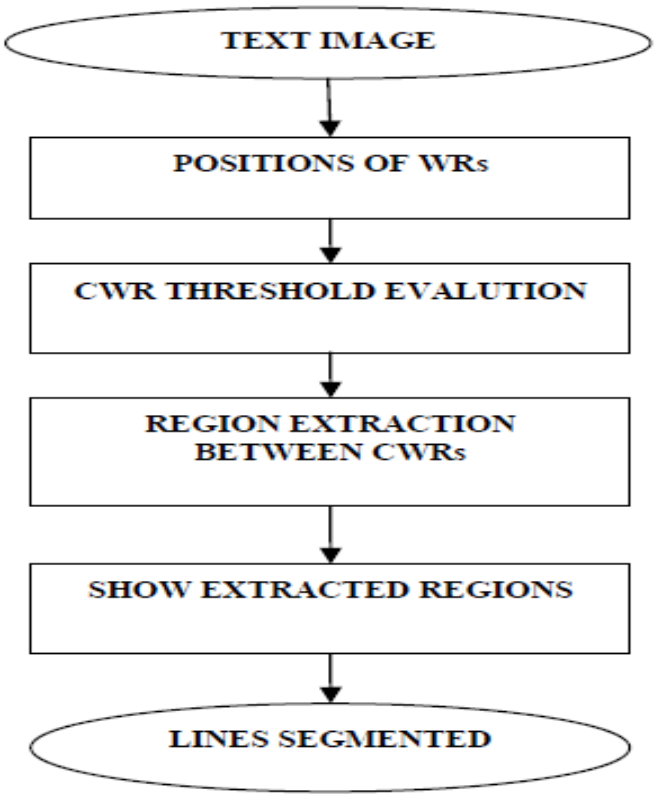


Figure.3 Design flow of line Segmentation method

Word Segmentation:

Subsequent operation of segmentation of a text line can be achieved with the tool to segment each words.In the line segmentation method we have already found the exact position of the reference text line that is the starting and ending rows in between which the text line exists. In this process we are finding a column of the image matrix that is common to the rows present between starting and ending rows of this text line and must have white pixel elements for all the concerned rows. We call this kind of column as a white column(WC).This process for finding WCs is continued throughout the text line till we bump into the last column of each concerned row. Thus we have located the positions of each words present within a text line and the position of white columns can guide the MATLAB processor to segment each words efficiently abandoning the in-between spaces.

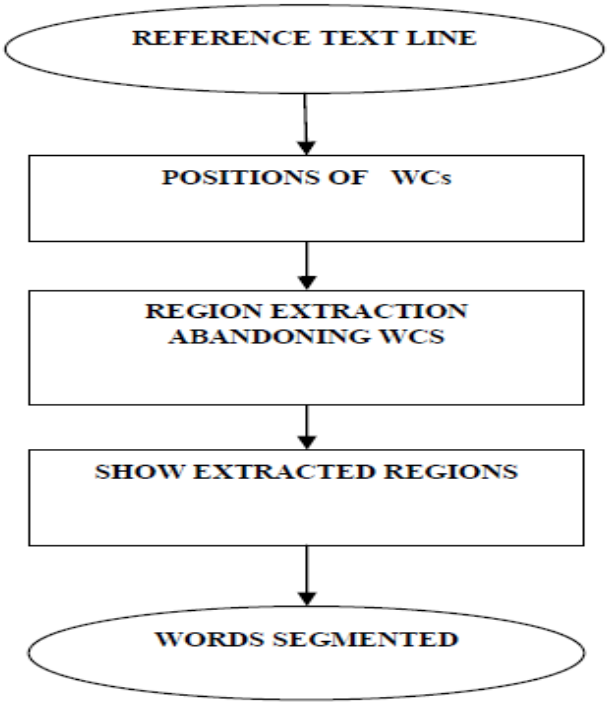
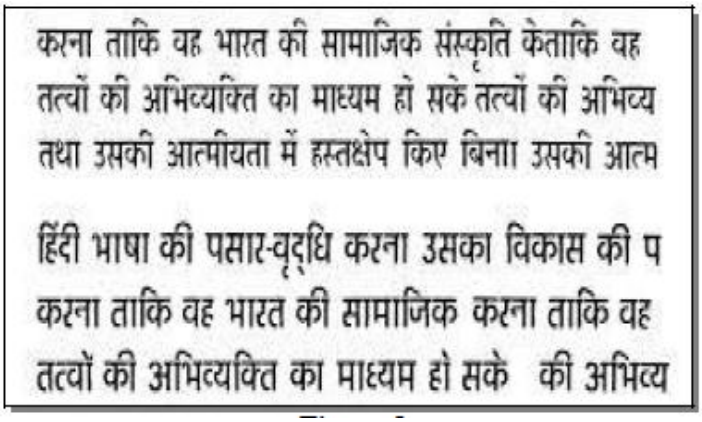


Figure.4 Design flow of Word Segmentation method

V. Results And Discussion

For performing the segmentation process ten different text images had been taken as input for this technique. It was found to be efficient than other techniques with 100% accuracy for line segmentation and 100% accuracy for word segmentation. Figure .5 act as an image which is taken as reference image for segmentation process. The table.1describes the performance comparison between the proposed work results and the previous results. The text line in figure.6.d is taken as reference in this paper to explain about the manners in which results are obtained. The below figure.6a to 6f describes the Line segmentation results and the figure.7 describes the results of word segmentation process.


करना ताकि वह भारत की सामाजिक संस्कृति केताकि वह
तत्वों की अभिव्यक्ति का माध्यम हो सके तत्वों की अभिव्य
तथा उसकी आत्मीयता में हस्तक्षेप किए बिना उसकी आत्म

हिंदी भाषा की पसार-वृद्धि करना उसका विकास की प
करना ताकि वह भारत की सामाजिक करना ताकि वह
तत्वों की अभिव्यक्ति का माध्यम हो सके की अभिव्य

Figure.5


करना ताकि वह भारत की सामाजिक करना ताकि वह

Figure.6.a

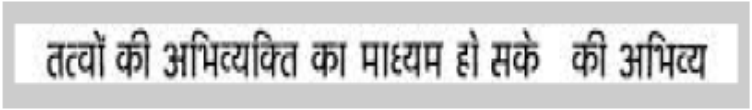

तत्वों की अभिव्यक्ति का माध्यम हो सके की अभिव्य

Figure.6.b

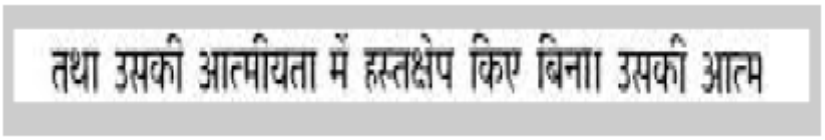


Figure.6.c

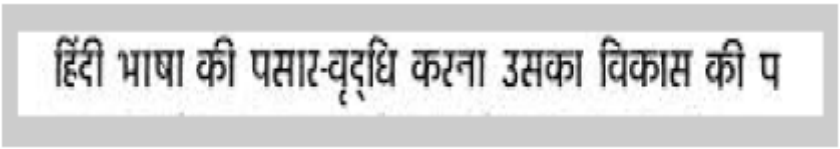


Figure.6.d

Figure.6.e

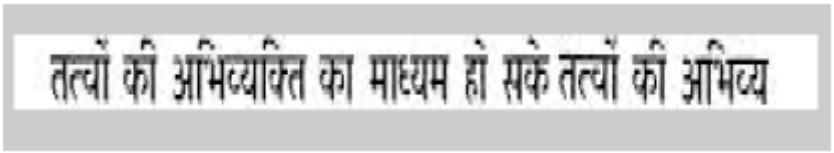


Figure.6.f

Figure.7

	Number of Lines	Number of Words	Lines Segmented Accuracy	Words Segmented Accuracy	% of Accuracy in Line Segmentation	% of Accuracy in Word Segmentation
[1]	8	45	8	42	100%	91%
[2]	200	1380	183	1354	91.5%	98.5%
[4]	10	---------	10	----------	100%	----------
[9]	----------	10	----------	10	--------	100%
This Work	6	55	6	55	100%	100%

Table.1 Performance Comparisons

VI. Colclusions

In this paper, we have presented an Efficient line and word segmentation technique which is very different from conventional methods that are being used.The efficient segmentation technique for machine printed Devanagiri script for both line and word segmentation process is simply based on recognition of spaces that separates two lines or two words and we found that the experimental results with different scripts are excellent with 100% accuracy for both line and word segments. The Performance of the proposed method is compared to the State-of-the-art technique. The experimental results have been presented which shows that the proposed technique is more superior to the other forms of the technique. Hence our method out performs its counterparts. This method is 100% efficient for machine printed Devanagiri scripts (for both line and word segmentation).

References

[1]. Vikash J Dhongre, Vijay H Mankar, "International Journal Of Computer Science, Engineering and information Technology(IJCSEIT)",Vol 1,No.3,August 2011

[2]. Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal, "International journal of computer Applications(0975-8887)",Volume 1-No.-4,2010

[3]. Nallapareddy Priyanka, Srikanta Pal, Ranju Mandal, (2010) "Line and Word Segmentation Approach for Printed Documents", IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition-RTIPPR, pp 30-36

[4]. Raghuraj Singh, C. S. Yadav, Prabhat Verma, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network", International Journal of Computer Science & Communication, 2010.

[5]. M. K. Jindal, R. K. Sharma and G. S. Lehal, "Structural Features for Recognizing Degraded Printed Gurmukhi Script", in Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008), pp. 668-673, April 2008.

[6]. V. H. Mankar et al, (2010) "Contour Detection and Recovery through Bio-Medical watermarking for Telediagnosis", International Journal of Tomography & Statistics, Vol. 14 (Special Volume), Number S10.

[7]. G. Louloudis, B. Gatos, I. Pratikakis and K. Halatsis, "A Block Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", in the proceedings of Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, pp. 515-520, 2006.

[8]. L. Likforman-Sulem and C. Faure, "Extracting text lines in handwritten documents by perceptual grouping", Advances in handwriting and drawing : a multidisciplinary approach, C. Faure, P. Keuss, G. Lorette and A. Winter Eds, Europia,Paris, pp. 117-135, 1994.

[9]. I.S.I. Abuhaiba, S. Datta and M. J. J. Holt, "Line Extraction and Stroke Ordering of Text Pages", in the Proceedings of Third International Conference on Document Analysis and Recognition, Montreal, Canada, pp. 390-393, 1995.

[10]. C. Weliwitage, A. L. Harvey and A. B. Jennings ,"Handwritten Document Offline Text Line Segmentation", in the Proceedings of Digital Imaging Computing: Techniques and Applications, pp. 184- 187, 2005.

[11]. A. Zahour, B. Taconet, L. Likforman-Sulem and Wafa Boussellaa, "Overlapping and multi-touching text-line segmentation by Block Covering analysis", Pattern analysis and applications, 2008.

[12]. Veena Bansal, "Integrating knowledge sources in Devanagari text recognition", Ph.D. thesis, IIT Kanpur, INDIA, 1999

Authors Profile

Mr. Siba Kumar Panda was born on November 09,1989.He received the B.Tech degree in Electronics & Communication Engineering from Biju Patnaik University of Technology,odisha in 2012 and M.Tech degree in VLSI Signal Processing Specialization from Veer Surendra Sai University Of Technology,odisha in 2014.

Currently he is working as an Assistant Professor at Centurion University of Technology and Management, Bhubaneswar; Odisha.He also awarded the University Silvermedal for best Electronics & Telecommunication Engineering Post Graduate for the academic year 2012-2014 at VSSUT, Odisha.

His research area of interest includes Ultra-wideband (UWB) device design, RF circuit design using CMOS Technique,VLSI implementation of Vedic Mathematics,VLSI Signal Processing.

Ms. Smruti Snigdha Pani was born on May 21st 1990.She received the B.Tech degree in Electronics & Communication Engineering from Biju Patnaik University of Technology,odisha in 2011 and M.Tech degree in Communication System Engineering from Veer Surendra Sai University Of Technology,odisha in 2014.

Her research area of interest includes Digital image processing ,ImageCompression ,Image Segmentation and Signal Processing.

Mr. Biranchi Narayan Panda was born on 12/01/1992.He received the B.Tech degree in Electrical and Electronics Engineering from Biju Patnaik University of Technology,odisha in 2013. He was working as an project engineer at Council of Scientific and Industrial Research(CSIR-NAL) from 4th August 2014 to 19th Sept 2014.

His research area of interest includes Digital image processing ,Image Compression ,Image Segmentation, Signal Processing and Electrical power quality design.