ISSN:0975-9646

Priyanka Karmakar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 157-160

Line and Word Segmentation of a Printed Text Document

Priyanka Karmakar, Biswajit Nayak, Nilamani Bhoi

Department of Electronics and Telecommunication Engineering
Veer surendra Sai University of technology, Burla, odisha, India

***Abstract*— Segmenting accurately a script document to extract various features that the document contains is a very challenging work and a need concerted effort. Continuous research works are in field to make the segmenting process simple and efficient. A simple segmenting technique for a line and word segmentation of a script document has been proposed. In this space recognition technique the main objective is to recognize the spaces that separate two text lines and the similar procedure is followed for the word segmentation procedure. Three different scanned document have been taken as input images for line and word segmentation experiment and result found were promising, with average accuracy for line and word were, 100% for line segmentation and 100% for line segmentation as well.**

Key words: shirorekha, white rows, white columns, consecutive white rows(CWR) threshold

I. INTRODUCTION

Importance of segmentation technique accumulating periodically at its practical application base is expanding rapidly. It is the primary stage for numerous processes such as machine recognition of language script. Segmentation is also used to extract various useful features of a document. One such 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 concentrates upon segmentation of Devnagiri script, which is an Indian script for writing number of languages like Hindi, Marathi, Sanskrit, Sindhi and Nepali languages.

RELATED WORKS

Continuous research works are in field to develop simpler and efficient techniques for line and word segmentation. Some important proposed techniques are thinning approach based segmentation[5], segmentation using histogram approach[1], header line and base line detection based method [2] , Hough transform based method [6], smearing method [7], grouping method [8], graph based method [9], CTM (Cut text Minimum) approach [10],
Block covering method [11] ,text line identification[12]. Each method has its own pros and cons. Some methods are briefed below:

Histogram approach

This method is based on pixel histogram obtain. Here a Y-histogram projection is performed which results in text line position. To divide a line into different regions a threshold is applied. After that another threshold is used to eliminate false lines. These procedures however, cause some loss on the text line area. So recovery method is proposed to minimize the effect.

Header line and base line detection based method

In this method we calculate the header line and base line of a text document for line segmentation. Header lines are rows with maximum number of black pixel and base line are rows with minimum number of black pixel. Till now researchers are detecting the header line by detecting finding the row with maximum pixel density, but it cannot work for skew variable test.

Hough transform approach

This approach enables us to detect collinear edge pixel even though each of them is isolated. This approach is useful to find lines in noisy images where local information around each edge pixel is unreliable or unavailable. Some problem in this approach is that it requires a relatively large amount of memory and a long computation time and raises the so-called connectivity problem, when illusionary lines composed of accidentally collinear edge pixel are also detected.

Smearing method

Run length smoothing method algorithm is a smearing method. In smearing method, 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.

Grouping approach

Elementary line segments(ELS's) 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.

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.
The paper is organized as follows: section II contains characteristics of Devnagiri script. Section III shows research works done upon segmentation. Section IV narrates the proposed method. Section V details the experimental results and conclusions are discussed in Section VI.

II. DEVNAGIRI SCRIPT

Devnagiri is an ancient Indian script that is widespreadly used to write some of the popular Indian languages. Hindi (written in Devnagiri script) is the national language of India and third most popular language in the world[3]. It is used as oral language by more than 500million people[3].
Hindi is written from left to right. Alphabets of Hindi contains 11 vowels(figure.2) and 33 consonant(figure.1). An uniqueness of Devnagiri script is presence of Shirorekha. Shirorekha is drawn over all the characters of a concerned word. Detailed characteristics of Hindi language are explained in [4].
A Devnagiri text line can be classified into three different zones with reference to position of the header line. The zone residing over the header line is known as upper zone, the zone which hosts basic characters is known as middle zone, and the zone which contains vowels or modifiers is known as lower zone.

Figure.1

Figure.2

III. PROPOSED METHOD

The proposed 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. A design flow for line segmentation is given in figure.3 and for line segmentation is given in figure .4. This method works as follows:

LINE SEGMENTATION

The principle for line segmentation lies in finding the white pixels in each row of a given text document taken in jpg format. Foremost step to start with the line segmentation starts with the rows which has its all the elements as white pixels, i.e. the row which does not have any of its element with pixel value zero. We call this kind of row as a white row(WR). An design flow chart is given is figure.3.

Then appearances of the consecutive white rows(CWRs) through out the image are taken into consideration. A threshold value for the number of times for which the WRs appear serially is assumed. This threshold value(CWR threshold) determines the spaces present in between text lines of the concerned image document. Special consideration has to be paid towards upper zone and lower zone of the Devnagiri text line as some spaces separates these zones from the middle zone, so that they are not abandoned by the processor as white spaces.

Now we are ready to do the segmentation part with out any inconvenience. Utility of the assumed threshold number is that it will guide the MATLAB processor to segment the exact text parts by abandoning the in-between white spaces.

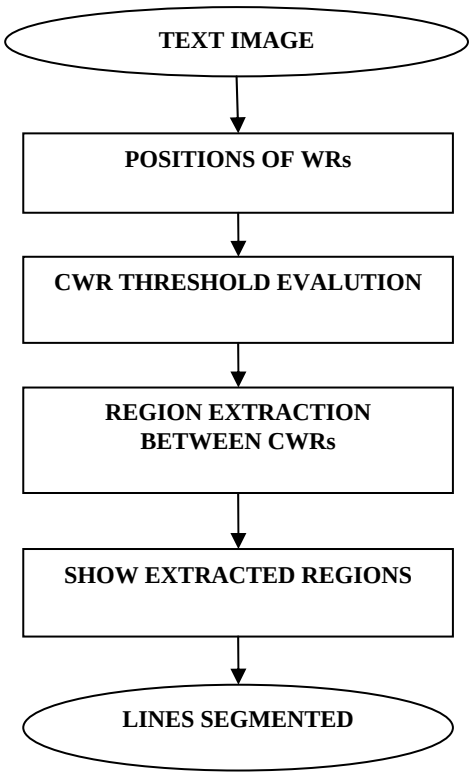

Figure.3

WORD SEGMENTATION

After successful segmentation of a text line we are ready with the tool to segment each words. 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. We find a column of the image matrix which 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 through out the text line till we encounter 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.

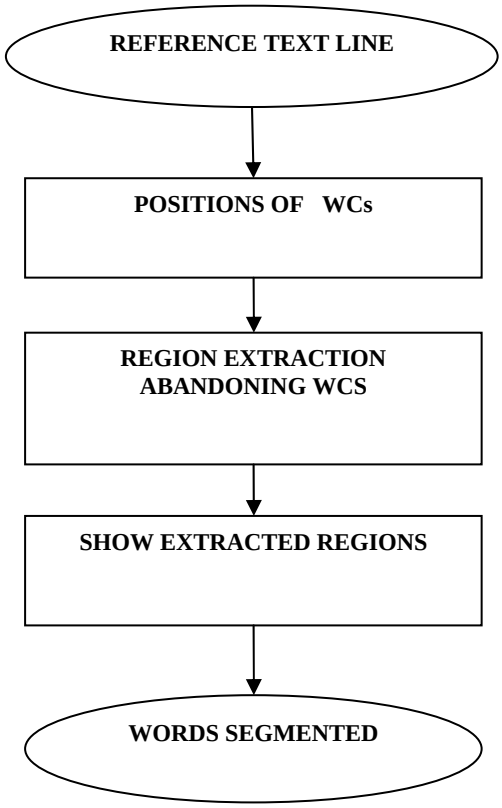

Figure.4

IV. EXPERIMENTS AND RESULTS

Ten different text images had been taken as input for this experiment and the technique was found to be efficient than other techniques with 100% accuracy for line segmentation and 100% accuracy for word segmentation. An image which is taken as reference for segmentation is given in figure(5). The following table.1 gives a comparison between my work results and the results given by Vikash J. Dhongre[1].

The text line in figure.6.d is taken as reference in this paper to explain about the manners in which results are obtained. Line segmentation results are shown in figure.6(a-f) and word segmentation results are shown in figure(7).

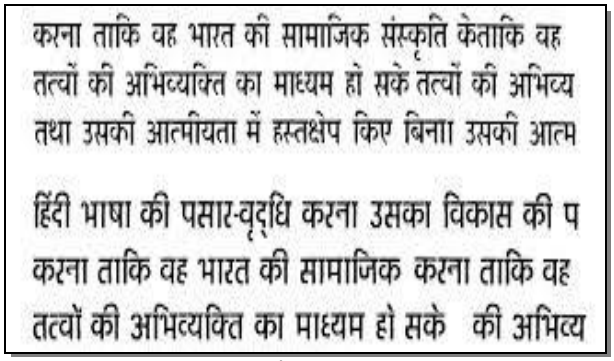

Figure.5

Figure.6.a

Figure.6.b

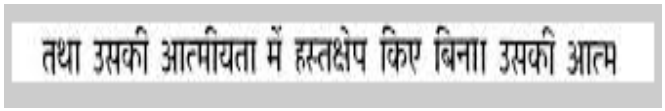

Figure.6.c

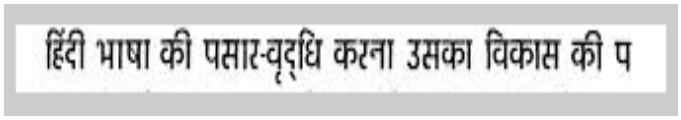

Figure.6.d

Figure.6.e

Figure.6.f

Figure.7

Table.1

By	No. of lines	No. of words	Lines segmented accurately	Words segmented accurately	Accuracy (line segmentation)	Accuracy (word segmentation)
Ours	6	55	6	55	100%	100%
Vikash J Dhongre	8	45	8	42	100%	91%

V. COLCLUSIONS

In this paper, we have presented a simple line and word segmentation technique which is very different from conventional methods that are being used currently like histogram based approach, projection based approach or thinning approach. The space recognition based approach proposed here is simply based on recognition of spaces that separates two lines or two words. And we found the experimental results with different scripts are excellent with 100% accuracy for both line and word segments.
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

Priyanka Karmakar received her B.tech from B.tech from Biju Pattnaik university of Technology, Of Technology,Rourkela, odisha in the year of 201year 2012 (Topper In her institute). Presently she presently she is pursuing her M.tech from Veer Sure from Veer Surendra Sai University of Technology, (Govt. Of Odisha University), Odisha. She has presented in various national and international conferences. Her research areas include image processing, antenna engineering and embedded systems.

Biswajit Nayak received his B.tech with first class with 1st class from Biju Pattnaik university of University of Technology Rourkela, odisha in the in the year of 2011. Presently he is continuing continuing his master degree University of from Veer Surendra Sai University of Technology Technohnology (Govt. Of Odisha University).Odisha. His research areas include image processing, wireless communication and embedded systems.

N. Bhoi received his B.E. in Electrical Engg. from University college of Engg., Burla, India and M.E. in Electronics and Telecommunication Engg. from Jadavpur University, Kolkata, He. completed his PhD in 2009 from National Institute of Technology, Rourkela, India. His research interest includes image processing and soft computing techniques. He is currently working as an associate professor in the department of electronics and telecommunication engineering at Veer Surendra Sai University of Technology, Burla, India