Offline Handwritten Devanagari Script Segmentation

Ashwin S Ramteke, Milind E Rane

Abstract— The process of segmentation is a vital part in any script/character recognition technique. Devanagari is mostly useful Script in India for number of officials and banking applications. Segmentation of Devanagari script is difficult because of presence of large character set which include vowels, consonants, compound characters and modifiers. This paper focus on the line, word, character segmentation of handwritten Devanagari script for efficient script recognition.

Index Terms— Handwritten Devanagari Script, binarization, line segmentation, word segmentation, character segmentation, connected component, projection profile.

1.INTRODUCTION

Segmentation of handwritten script is very important task for post processing of handwritten script recognition. Segmentation is important to improve the accuracy of handwritten script identification, since recognition system is heavily depends upon segmentation phase. Segmentation means to subdivide a handwritten script image into a particular part such as line, word or character. Basically in the segmentation approach it has been tries to extract a specific part of handwritten Devanagari script document images. The large variation in handwriting style of the script makes the task of segmentation quite difficult.

2 LITERATURE REVIEW

The process of segmentation having immense importance in the handwritten script recognition. Therefore an ample study of research outcome in related segmentation field was surveyed. A method of contour code feature based segmentation has proposed by Brijesh Verma for handwriting recognition ,where he used rule based segmentation for the improvement of handwriting recognition[1]. A segmentation approach proposed by Singh B, Gupta N,Tyagi R, Mittal A, Ghosh D, uses profiling based method which uses the vertical and horizontal density of black pixels along an axis[2]. In [3] Das, Reddy, Govardhan, Saikrishna proposed algorithm based on connected components and projection file. The segmentation algorithm in [4] proposed by Xia Liu and Zhixin Shi uses chain code representation character stroke and segmentation is done by splitting and grouping the contours. A segmentation technique over overlapping line of uniform size text on non headline based for distorted Tamil scripts done by Gandhi R.I, and Dr. Iyakutti K [5]. A Support Vector Machine based algorithm for segmentation of handwritten Devanagari script is proposed by Agrwal G, Mukerjee A, Kumar N in [6]. A method for line segmentation of handwritten Hindi text has proposed by Garg N.K., Kaur L., Jindal M.K. in [7]. The work proposed in this paper is modified version of their previous proposed method with some assumptions related with the consonant height, maximum height of consonant and lower modifiers and skew between two lines in a text. Pal and Dutta [8], also used the stripe based partial projection based method with water reservoir concept for line segmentation of Unconstrained Bangla Handwritten text. Water reservoirs are used to determine the height of the character. The method of graph cut proposed by Jawahar, Kumar, Namboodiri [9] requires a priori information about the script structure to cut. Shoba,Rajasekharan Sagar propped a method based on projection method for Kannada script document segmentation [10]. In projection based method horizontal projections of the pixels are used to segment the text into lines. This method is suited for straight lines or easily separable lines only. This method is modified by some researchers using partial projection method [11]. The input text is divided into vertical stripes and horizontal projection of each stripe is considered for line segmentation. In smearing method, fuzzy run length is used for line segmentation. The fuzzy run length describes how far one can see when standing at a pixel along horizontal direction. In [12], the smearing method is used for text line segmentation. Run length smearing algorithm is used to segment individual text lines from document images. The threshold for RLSA is computed based on the height information of the text lines. In ICDAR 2009 Handwriting Segmentation Contest various methods on handwriting segmentation techniques proposed by various authors contributed by Gatos B. Stamatopoulos N., Louloudis G. [13]. A printed Devanagari script segmentation algorithm for line segmentation based on horizontal histogram observation , for word segmentation defining the boundary for word and for character segmentation extracting of the word done in [14] by Shukla M,K., Patnaik Tushar, Tiwari S., Dr. Singh S.K. As Segmentation into isolated character has played a crucial step in for handwriting recognition system. Therefore, the horizontal projection of a document image is the most commonly used technique to extract the lines from the document [15-18]. Thus, segmentation plays an important role in improving the overall process of handwriting recognition. Various algorithms and approaches used by different researchers are discussed herewith in the form of review.

- *Ashwin S Ramteke is currently pursuing masters degree program in electronics and telecommunication with signal processing specialization in University of Pune, India, PH-07709187254. Email: ashwinramteke1@gmail.com*
- *Milind E Rane received his BE degree in Electronics engineering and M Tech in Digital Electronics from Visvesvaraya Technological University, Belgaum, in 1999 and 2001 respectively. His research interest includes image processing, pattern recognition and Biometrics recognition. Currently working as an Asst.Prof. at VIT Pune,University of Pune , India , PH-09545456645. E-mail: me_rane@rediffmail.com*

Now before starting to discuss about the approach used for the segmentation in this paper, initially it would be better to discuss some of the characteristics of Devanagari script.

3 CHARACTERISTICS OF DEVANAGARI

Devanagari originated from ancient Brahmi script through various transformations. As there is typically a letter for each of the phonemes in Devanagari, the alphabet set tends to be quite large. Devanagari script has 13 vowels, 34 consonants, 14 modifiers of vowels and of rakars. Also Devanagari having compound characters which are formed by combining two or more basic characters. It is a phonetic and syllabic script, words are written exactly as they are pronounced. Apart from the above features another distinctive feature of Devanagari is the presence of a horizontal line on the top of all characters. This line is known as header line or maatra. It also contains the ten numerals whose combinations in different way can form large set of numbers used to describe amount. Following figure 1 will show example of the Devanagari script.

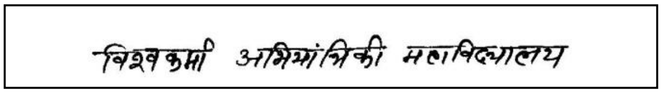


Figure 1 Example of Devanagari Script

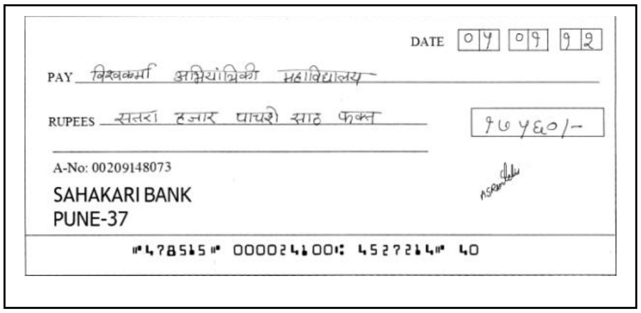
4 PROPOSED SEGMENTATION METHODOLOGY

The process of segmentation includes separating word, line, individual character or pseudocharacter images from a given script image. The large variation in handwriting style and of the script makes the task of segmentation quite difficult. Before proceeding to the process of segmentation preprocessing is need to be done. In the preprocessing smoothing of image using median filter and the binarization of image and scaling is include. In this approach of segmentation, the header line is present in the input script. The process of segmentation consists in analyzing the digitalized image provided by a scanning device, so as to localize the limits of each character and to isolate them from each other. In the handwritten Devanagari script the space between the words and the characters may varies which may produce some difficulties in the segmentation process of handwritten Devanagari script. For the line segmentation connected component approach is used. Clustering the connected components to extract the line. For the segmentation of the handwritten Devanagari script into words, vertical projection profile i.e. the histogram of input image, where the zero valley peaks shows the space between the words and characters. Finding the maximum character space and used it for separating the words. For the character segmentation of Devanagari, using the vertical profile to separate the base character using clear paths between them [3]. The steps of proposed algorithm for segmentation of handwritten Devanagari script is follows. The algorithm is simulated in Matlab 7.0

Steps of Algorithm

```
{
Start
Read image                %%  for number for required lines
Crop image                %%  to process from input image
Check the image           %%  if white background
background                %%  convert it to black
check the boundry         %%  to add background to image
apply line extract function
apply word extract function
apply character extract function
show output
end
}
```

The result obtained from this algorithm are as shown below. The results are apply on the format of input bank cheque image, as shown in Figure 2



DATE ०४ ०९ ९२

PAY विश्वकर्मा अभियांत्रिकी महाविद्यालय

RUPEES सतरा हजार पाचशे साठ फक्त

१७५६०/-

A-No: 00209148073

SAHAKARI BANK
PUNE-37

⑈478515⑈ 000024100⑆ 4527214⑈ 40

figure 2 example input bank cheque image

figure 3 (a) will show the segmented output of input image of pay name script in to words and figure 3 (b) will show the segmented character output for given script in Figure 2.

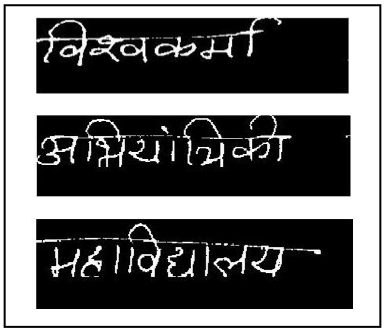


figure 3 (a) segmented words from script

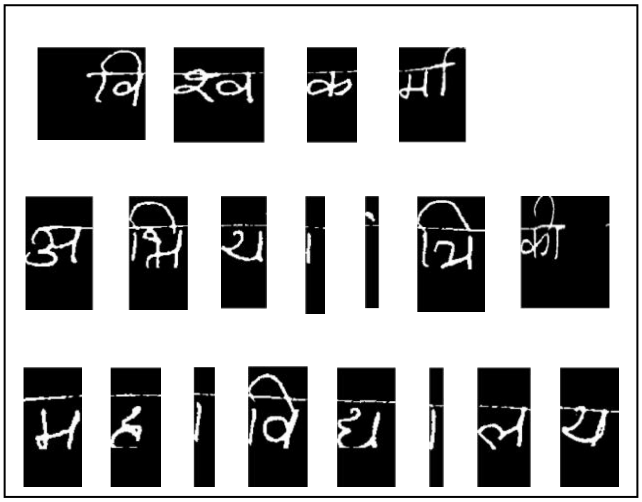


figure 3 (b) segmented characters from words

Following figure 4 will show the input script for rupees and segmented output of input script image into words and characters.

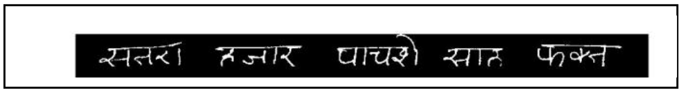


Figure 4 (a) input script image for rupees

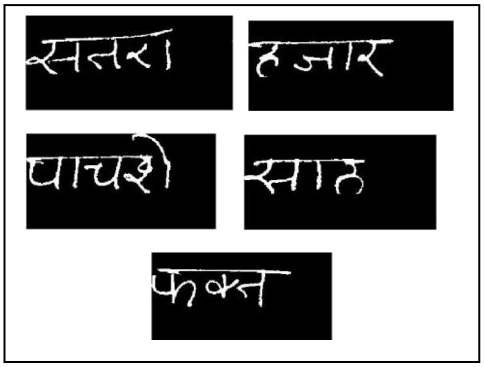


Figure 4 (b) segmented script output in to words

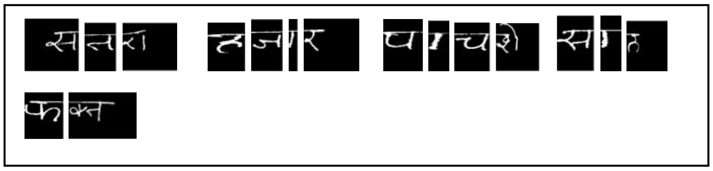


figure 4 (c) segmented word output in to characters

Figure 5 will show the input image of amount in number and segmented output.

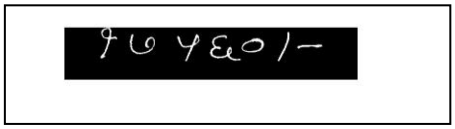


figure 5 (a) input amount image in numbers

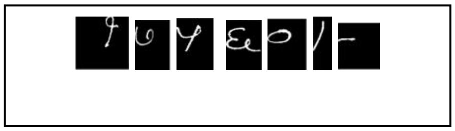


figure 5 (b) segmented output for input figure 5(a)

5 Conclussion

The handwritten data set are collected from different users of different background on blank bank cheque format as well as on plane papers. The algorithm is implemented in the Matlab. The algorithm is tested with the large number of input images. The segmentation accuracy for this implementation depends upon the proper writing i.e. non-overlapping or characters, proper space between words and characters, proper connection of characters through shirorekha. The segmentation for word gives 98% of accuracy, for characters 97% of accuracy. The implementation not gives that much accurate result for the broken characters. For numerical segmentation implementation gives the 100% accurate result.

References

[1] Brijesh Verma, “A Contour Code Feature Based Segmentation for Handwriting Recognition”, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).

[2] Brijmohan Singh, Nitin Gupta, Rashi Tyagi, Ankush Mittal, Debashish Ghosh, “Parallel Implementation of Devanagari Text Line and Word Segmentation Approach on GPU” , *International Journal of Computer Applications (0975 – 8887) Volume 24– No.9, June 2011*

[3] Das, Reddy, Govardhan, Saikrishna, “Segmentation of overlapping text lines, characters in printed telugu text document images” International Journal of Engineering Science and Technology Vol. 2 (11), 2010, 6606-6610. R. Nicole, "The Last Word on Decision Theory," *J. Computer Vision,* submitted for publication. (Pending publication)

[4] Xia Liu and Zhixin Shi, “A Format-Driven Handwritten Word Recognition System**”,** Proceedings Seventh International conference on Document analysis and recognition , 2003, page 1118-1122

[5] Gandhi R. I., Dr. Iyakutti K, “A Technique for Segmentation over Overlapping Line of Uniform Sized Text on Non-Headline Based Distorted Tamil Scripts”. Int. J. of Advanced Networking and Applications, Volume: 02, Issue: 02, Pages: 491-495 (2010).

[6] Agrawal G, Kshitij, Mukharjee A, Kumar N, “Handwritten Devanagari Script Segmentation using Support Vector Machines” International conference on Neural Network (IJCNN), 2004.

[7] Garg N.K., Kaur L., Jindal M.K., “A New Method for Line Segmentation of Handwritten Hindi Text**”** Seventh International Conference on Information Technology, 2010.

[8] U. Pal and S. Datta, “Segmentation of Bangla Unconstrained Handwritten Text”, Proceedings of the 7th International Conference, ICDAR, pp.1128-1132, 2003

[9] K.S. Sesh Kumar, A. M. Namboodiri, C.V. Jawahar. (2006): Learning Segmentation of Documents with Complex Scripts, Fifth Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, India, LNCS 4338, pp.749-760.

[10] B.M. Sagar, DR. G. Shoba, DR. P. Ramakanth Kumar. (2008): “Character Segmentation algorithms for Kannada optical character.” International conference on Wavelet analysis and Pattern recognition, 2008. page 339-342..

[11] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, “Arabic Hand-written Text-line Extraction”, Proceedings of the Sixth International Conference on Document analysis and recognition, 2001, page 281-285

[12] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, “A new algorithm for detecting text line in handwritten documents”, Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, pp. 35–40, 2006.

[13] Gatos B. Stamatopoulos N., Louloudis G., “ICDAR2009 Handwriting Segmentation Contest” 10th International Conference on Document Analysis and Recognition, 2009.

[14] Shukla M,K., Patnaik Tushar, Tiwari S., Dr. Singh S.K., “Script Segmentation of Printed Devnagari and Bangla Languages Document Images OCR” International Journal of Computer Science and Technology vol 2, issue2, June 2011.

[15] U. Pal and B. B. Chaudhuri, .Printed Devanagari script OCR system., Vivek, Vol 10(1), pp. 12-24, 1997.

[16] B. B. Chaudhuri and U. Pal, .A complete printed Bangla OCRsystem., Pattern Recognition, Vol. 31(5), pp. 531-549, 1998.

[17] G. S. Lehal, C. Singh and R. Lehal, .A shape based post processor for Gurmukhi OCR., in the Proceedings of 6th ICDAR, pp. 1105-1109, 2001.

[18] A. Goyal, G. S. Lehal and S. S. Deol, .Segmentation of machine printed Gurmukhi script., in the Proceedings of 9th International Graphonomics Society Conference, Singapore, pp. 293-297, 1999.