An Efficient Segmentation Scheme for the Recognition of Printed Devanagari Script

[1]Manoj Kumar Shukla, [2]Dr. Haider Banka
[1]Dept. of CSE, Sunder Deep Engineering College, Ghaziabad, UP, India
[2]Dept. of CSE, India School of Mines, India

Abstract
In this paper we propose on the script segmentation of Devanagari character for efficient recognition of the character. Script segmentation is a very important step in the process of the recognition a Devanagari script document or and other script document. Segmentation process is allows the OCR system to classify the Devanagari or other characters to improve the accuracy of the script Line or character. At the time of script or character recognition we have found that incorrect segmentation leads to incorrect recognition. Segmentation phase includes line, word and character segmentation. At the time of Recognition or Identification of the Devanagari script, we have found a difficult problem to segment the Devanagari script. Devanagari characters or scripts, lines are touching with head line. Some character shapes are similar. There are above 300 compounds, modified and basic shapes. Hence in the Devanagari script segmentation is a very challenging task at the time of recognizing the character of script document. In this paper we have achieved above 99.5% average accuracy depending upon Devanagari script document.

Keywords
Optical Character Recognition, Pre-Processing, Devnagari Script.

I. Introduction

India is a multi script, multi lingual country. So in India researchers have developed a successfully multi-lingual OCR system. So a very important step to identify the Indian language script before feeding each language script line document in OCR system. In India there are 23 official languages like, Devanagari, Bangla, Gurumukhi, Sanskrit, Tamil, Telugu, Malayalam, Maithili, Nepali, Oriya, Assamese, Santali, Sindhi, Bodo, Manipuri, Marathi, English, Dogri, Guajarati, Kannada, Kashmiri, Konkani, and Urdu. There are 13 different scripts like (Devanagari, Bangla, Gurumukhi, Kannada, Malayalam, Oriya, Roman, Gujarati, Assames, and Kashmiri. Optical Character Recognition (OCR).OCR is a process of automated reading of the text line from scanned image. OCR is the most challenging research field in Document processing, Image processing and pattern recognition. Devanagari is the most common used language and script in India. It is also called Hindi. Devanagari is the national language in the India. It is third most popular language in the word. At the time of script recognition we have found in OCR system a large number of recognition error due to character segmentation. Devanagari script segmentation has been a challenging research problem at the time of recognition the script in OCR system. Our investigation we have found some work in Devanagari script segmentation. Manoj et al [1] have used techniques for segmentation the two Indian language script. Dhalakea et al [2] suggested a method for stapes of connected component to find the three zones in the printed Gujarati script. Chaudhuri et al [3] have proposed a very useful method structure and statistical features for separating machine printed text line for hand – written text line for both Bangla and Devnagari scripts. Harikumar et al [4] describe the concept of average line high to segment the Malayalam script. Pal et al [5-6] have used the concept of zoning and line segmentation. Singh et al [7-8] constructed the simple method of line, word and character segmentation. Bansal et al [9] have used a two pass algorithm based upon average line height to solved the segmentation problem in Devanagari script.

II. Features of Devnagari Script

Devanagari is the script for Hindi which is official language of India. It is also the script for Sanskrit Marathi, and Nepali languages. The script is used by more than 450 million people on the globe. Devnagari script is a logical composition of its constituent symbols in two dimensions. It is an alphabetic script. [1]. During survey we have observed that the name Devnagari came from the Sanskrit word Diva (god), and Negara (city); In the other sense we can say the script of the city of the God. Devanagari has 11 vowels and 33 simple consonants. These are called basic characters. In Devanagari script we have found that besides the consonants and the vowels, other constituent symbols in Devanagari are set of vowel modifiers called matra (placed to the left, right, above, or at the bottom of a character or conjunct), pure-consonant (also called half-letters) which when combined with other consonants yield conjuncts. In a Devnagari character a horizontal line called shirorekha (a headerline) runs through the entire span of work. Our literature of the Devanagari script it is a derivative of ancient Brahmi script which is mother of almost all Indian scripts. Word formation in Indian scripts follows a definite script composition rule for which there is no coun- terpart in Roman. At the time of recognition for the Devanagari script we have found similar shape. That's why Devanagari character recognition or segmentation is a very challenging task in Optical Character Recognition. Here is simple Devanagari character property [10].

Table 1: Vowels and Corresponding Modifiers

Vowels:	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Modifiers:		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

Table 2: Consonants

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट
ठ	ड	ढ	ण	त	थ	द	ध	न	प	फ
ब	भ	म	य	र	ल	व	श	ष	स	ह

Table 3: Half Form of Consonants with Vertical Bar

क्	ख्	ग्	घ्		च्		ज्	झ्	ञ्	
			ण्	त्	थ्		ध्	न्	प्	फ्
ब्	भ्	म्	य्		ल्	व्	श्	ष्	स्	

Table 4: Examples of Combination of Half-Consonant and Consonant

क क्क	क ल्क्ल	घ न घ्न	च ञ च्ञ	ञ च ञ्च	त न त्न	प त प्त	प ल प्ल
व व व्व	भ न भ्न	म ल म्ल	ल ल ल्ल	ञ न ञ्न	ञ व ञ्व	ञ ल ञ्ल	म न म्न

Table 5: Examples of Special Combination of Half-Consonant and Consonant

क ष क्ष	ज ञ ज्ञ	ट ट ट्ट	ट ठ ट्ठ	त र त्र	द द द्द
द ध द्ध	द व द्व	द व र द्व	श र श्र	द भ द्भ	द य द्य

Table 6: Special Symbols

क़	ख़	ग़	ज़	फ़	ड़	ढ़	ँ	ॅ	ः	ऽ	ॅ

III. Pre-Processing

In my previous paper [11] define Pre-Processing steps for Indian languages script document. Pre-processing is the name given to a family of procedures for smoothing, enhancing, filtering, cleaning-up and otherwise massaging a digital image so that subsequent algorithms along the road to final classification can be made simple and more accurate. In general we are using scanned document images for script identification or OCR system where document images are not good candidate for segmentation and identification of script. Pre-processing step are used after scanning the document. The preprocessing steps for Binarization, background noise cleaning for scanning images, skew correction of the scanned images are given below.

A. Binarization or Text Digitization

Binarization is an important step of script identification for OCR (Optical Character Recognition) system. Binarization step is mainly applied for scanned images. In this step original scanned images convert into two tone images. The two tone images are converted into 0 and 1 label that means the 0 represents the background of the images or also can say white area of the scanned images and 1 represents the pixel area of the scanned images document or we can say object area of the scanned document images.

B. Noise Removing or Cleaning

Pre-processing steps is used in script identification or OCR system to improve accuracy of the result. Generally we are using scanned document images for script identification or OCR system because document images are not of good quality it is also generating noise. In the present work the noise is removed by taking a threshold of the pixel value of the grey image of the image document. In grey scale images the pixel values lies between 0-255. It has been found from the observations of many document images that the pixel which actually belongs to any character has minimum value ranging from 40-50. It is also found that the pixels having values outside the range 40-50 carrying no information about the handwritten characters. In our observation we find that for most of the document images the noisy pixels have values greater than 200.

C. Skew detection and corrections

Document originally has zero skew, but when a page is manually scanned or photocopied, nonzero skew may be introduced. The skew may cause problems in text baseline extraction and document layout analysis techniques which assume the Manhattan layouts, that is, the layouts whose blocks are separable by vertical and horizontal cuts. Therefore, it is often necessary to determine the skew angle before structural analysis.

IV. Segmentation of printed Devnagari and Bangla script

Script segmentation of a document images is a very important task for a character recognition system. Before segmentation step we are using very important step preprocessing (i.e.- skew detection, Noise removed, binarization, Normalization and thinning).Script segmentation for Bangla and Devnagari script are more complex. Script segmentation is done by following process.

A. Line Segmentation

Line based script segmentation method all black pixels on every row are computed horizontally. In this method we are separating individual line in a script document image based on the peak of the horizontal Histogram. In this step i.e. line based script segmentation we are constructs the horizontal histogram for the image. Results are given below.

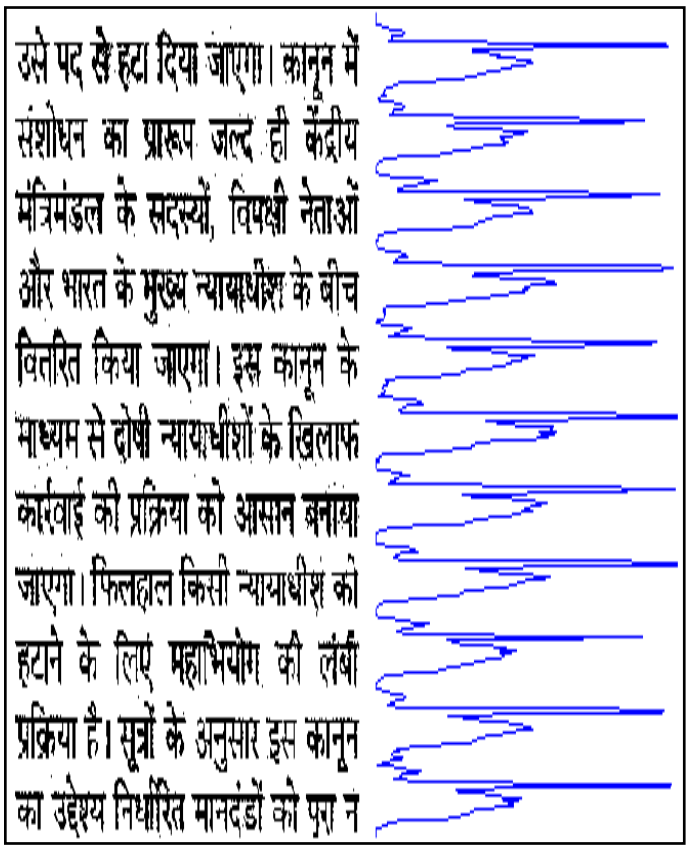

Fig. 1: Line based script segmentation for Devnagari script

B. Word Segmentation

Word segmentation means that we are separating individual word in a script document image based on the boundary of each word. In this step i.e. word based script segmentation we construct the boundaries for the image. Results are given below.

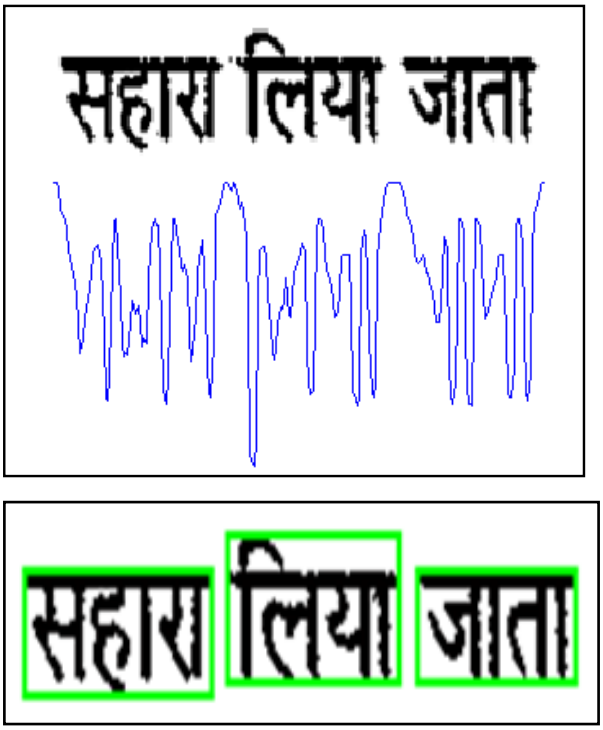

Fig. 2: Word based script segmentation for Devnagari script

C. Character Segmentation

Character segmentation means that we are separating individual character in a script document image based on the extracting of the word. In this step i.e. character based script segmentation we are separating each character for the image. Result is given below.

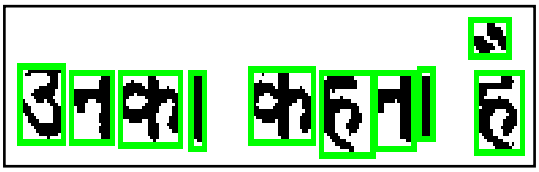

Fig. 3: Character based script segmentation for Devnagari script

V. Result

We applied our segmentation scheme on 500 different script line for segmentation line based approach, 500 data sample for word based segmentation approach and 500 data set for character based segmentation approach. From the experiment we have found 100% result of line based script segmentation, 100% of word based script segmentation and 99% character based segmentation. Result of different script segmentation techniques are shown in the following table.

Table 7:

Approach	Data Sample	Complete Segment	Rest	%
Line Based	500	500	0.00	100%
Word Based	500	500	0.00	100%
Character Based	500	495	5.00	99%

VI. Conclusions

In this paper we have presented a segmentation scheme for the Devanagari script that is very useful for the Image processing and Optical Character Recognition. At the time of recognition of the script proper segmentation of the script line or character helps improve the accuracy rate. This paper segmentation technique describes line based, word based, character based for Devanagari script language document. In the present algorithm we have achieved better performs of previous algorithm for line based script segmentation, word based script segmentation, and also character based script segmentation for Devnagari script document image.

References

[1] Manoj kr. Shukla, Tushar Patnaik Shrikant Tiwari, Sanjay Kumar Singh, "Script Segmentation of Printed Devnagari and Bangla Languages Document Images OCR", IJCST Vol. 2 Issue 2. 2011.

[2] J. Dholakia, A. Negi, S. R. Mohan, "Zone Identification in the Printed Gujarati Text", In Proceedings of the ICDAR (ICDAR'05), pp. 272-276, 2005.

[3] U. Pal, B. B. Chaudhuri, "Automatic Separation of Machine-Printed and Hand-Written Text Lines", In Proceedings of the ICDAR (ICDAR'99), pp. 645-648, 1999.

[4] S. Harikumar, K. Jithesh, K. G. Sulochana, R. Ravindra Kumar, "Script based line & character segmentation techniques for Malayalam document images", In Proceedings of the International Symposium on Machine Translation (iSTRANS 2004) New Delhi, India, pp. 122-127, 2004.

[5] B. B. Chaudhuri, U. Pal, "A complete printed Bangla OCR system", Pattern Recognition, vol. 31, no. 5, pp. 531-549, 1998,

[6] U. Pal, B. B. Chaudhuri, "Printed Devnagari script OCR system", Vivek, vol. 10, pp. 12-24, 1997.

[7] G. S. Lehal, Chandan Singh, "Text segmentation of machine printed Gurmukhi script", in Proceedings of SPIE, vol. 4307, pp. 223-231, 2001.

[8] G. S. Lehal, Chandan Singh,"A technique for segmentation of Gurmukhi text", In Proceedings of the 9th International Conference on Computer Analysis of Images and Patterns CAIP 2001, Warsaw, Poland, vol. 2124, pp. 191-200, 2001.

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

[10] Vikash J Dongre, Vijay H Mankar, " Devnagari Document Segmentation using Histogram Approach", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 1.1, No. 3, 2011.

[11] Manoj Kumar Shukla, Manoj Kumar Sharma,"Pre-Processing of Indian languages document images", International Journal of Engineering Science (Special issue 2011).