Line-wise Script Segmentation for Indian Language Documents

Manoj Kumar Shukla
Research Scholar
Department of CSE
ISM-Dhanbad

Haider Banka, Ph.D.
Associate Professor,
Department of CSE
ISM-Dhanbad

ABSTRACT

In a multi-lingual country like India, script segmentation or script separation of the multi-script in an image of a document page is of primary importance for a script identification system. For script segmentation of such a document page, it is necessary to segment multi script forms before running individual OCR of the script. In this paper we present a technique for script segmentation of the individual text line for printed in Indian language document. Our line wise script segmentation approach is Horizontal Projection Profile based script segmentation. A prototype of the system has been tested on printed Indian language lines of script and an average accuracy of 99% has been achieved.

Keywords

Script line documents, OCR

1. INTRODUCTION

India is a multi lingual, multi script country. Therefore developing a successful multi-lingual OCR, system for segmentation of different scripts is a very important step. In a multi-lingual country like India, it is very essential for designing an OCR system. In India 23 official languages namely Hindi, Punjabi Bengali, Maithili, Malayalam, Nepali, Oriya, , Sanskrit, Tamil, Telugu Assamese, Santali, Sindhi Bodo, Manipuri, Marathi Dogri, English, Gujarati, Kannada, Kashmiri, Konkani, and Urdu. There are 13 different scripts Gurumukhi, Devnagari, Bangla Assamese, Gujarati, Malayalam, Oriya, Roman, Kannada, Kashmiri. Optical Character Recognition (OCR) is the most Important and challenging area of Image Processing & pattern recognition useful in many practical applications like, reading aid for the blind, automatic reading for sorting of postal mail, bank cheques etc. For Optical Character Recognition (OCR) of such a document page, it is necessary to identify script forms before running individual OCR of the scripts. Although a lot of research work has been done for multi-script identification Indian languages script OCR technique have been developed over the years. Chaudhuri et al [1, 2, 3] have also suggested a method on separation of text line from different script using projection profile. Gaurav et al [4] have proposed a very useful method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bangla, Telugu and Urdu. Patil et al [5] proposed a method for neural network based system for script identification (Kannada, Devanagari, and English) document. V. Dhandra et al [6] describe word level script identification in bilingual document through discrimination features. Chanda et al [7] constructed an automatic technique for word wise identification of Devanagari, English and Urdu script for a single document. Padam et al [8] have proposed a method for English, Hindi and Kannad script identification using discriminating features and top and bottom profile based feature. Josh et al [9] have proposed a very useful method for script identification for Indian languages document. Zhou et al. [10] .

2. SEGMENTATION

Segmentation is a standout amongst the most imperative stages in character distinguishment handle. Segmentation is the methodology of portioning the entire record picture into unmistakable units for characteristic extractor and classifier. Message zone from the archive is concentrated and the division step is accompanied by portioning the content area into unique lines. Further, every line is sectioned into distinctive expressions, and at last, every saying is portioned into distinct characters. An info record might hold numerous sorts of data like photos, figures, various articles (perhaps in numerous segments) and so on. Three real parts of a complete content perusing framework are record dissection, report comprehension and character division [11]. The archive examination part extricates lines of content from a page for distinguishment. It likewise uncovers the constituents of an archive, for example photos, illustrations and content lines. The archive comprehension part removes legitimate relationships between the record constituents. The character division segment removes characters from a content line and recognizes them. Thusly, for an OCR framework, segmentation handle includes accompanying steps:

1. Detection of content locales in the info record.
2. Segmentation of content locale into unique lines.
3. Segmentation of content line into unique statements and zones.
4. Segmentation of word into unique character

(a) Horizontal Projection: For a given binary image of size$L×M$, where$L$is the heightand M is the width of the image, the horizontal projection is defined by Bansal [12] as: $HP(i)$,i= 1, 2, 3, …, L.

where$HP(i)$ is the total number of black pixels in i^{th} horizontal row.

(b) Vertical Projection: For a given binary image of size$L×M$, where$L$is the height and$M$is the width of the image, the vertical projection is defined by Bansal [12] as:

$VP(j)$, j = 1, 2, 3, …, M .

where$VP(j)$ is the total number of black pixels in j^{th} vertical column.

(c) Continuous Vertical Projection: For a given binary image of size$L$×M, where$L$is theheight and M is the width of the image, the continuous vertical projection has been defined as:

$CVP(k)$, $k = 1, 2, 3, \ldots, M$

Where $CVP(k)$ counts the first run of consecutive black pixels in k^{th} vertical column.

3. LINE WISE SCRIPT SEGMENTATION

For the most part, every content line is segmented from the past and emulating lines by white spaces. Along these lines, the flat projection of an archive picture is the most regularly utilized strategy to concentrate the lines from the report [13, 14, 15, 16]. Provided that the lines are decently differentiated and not tilted, the flat projection will have overall divided tops and valleys [77]. These valleys might be located effectively and used to verify the areas of the line verges. In debased printed Devnagari/Bangla script, applying the straightforward thought of flat projection to section the entire report into distinct lines does not work well. Over division happens when the white space breaks a content line into two or more level content strips as demonstrated in Figure 1 (issue ranges have been circled). In some cases, more level zone characters of one line touch the upper zone characters of next line, in this way generating on a level plane covering lines, called under division, as demonstrated in Figure 2 (issue regions have been enclosed). The issue of evenly covering lines is a regular issue in daily papers and magazines of printed Devnagari/Bangla script.

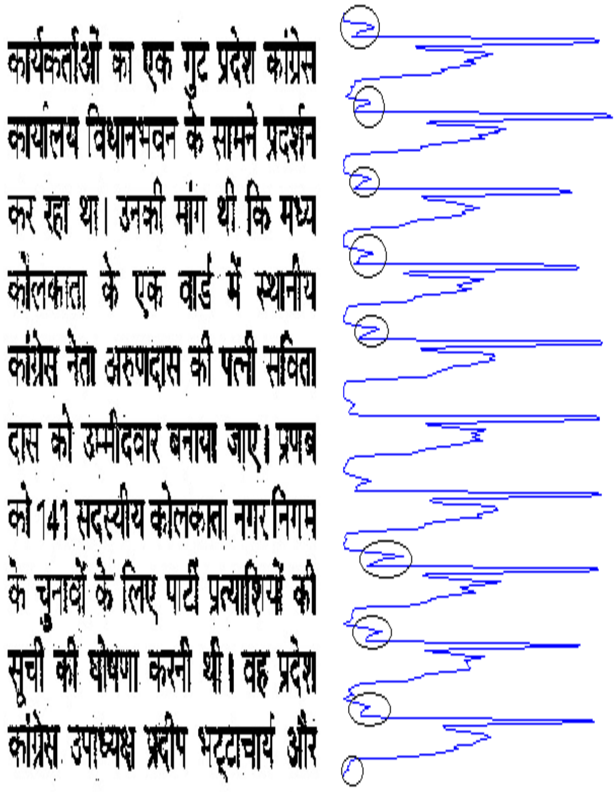

Figure 1: Horizontal projection of Devnagari script document resulting over segmentation

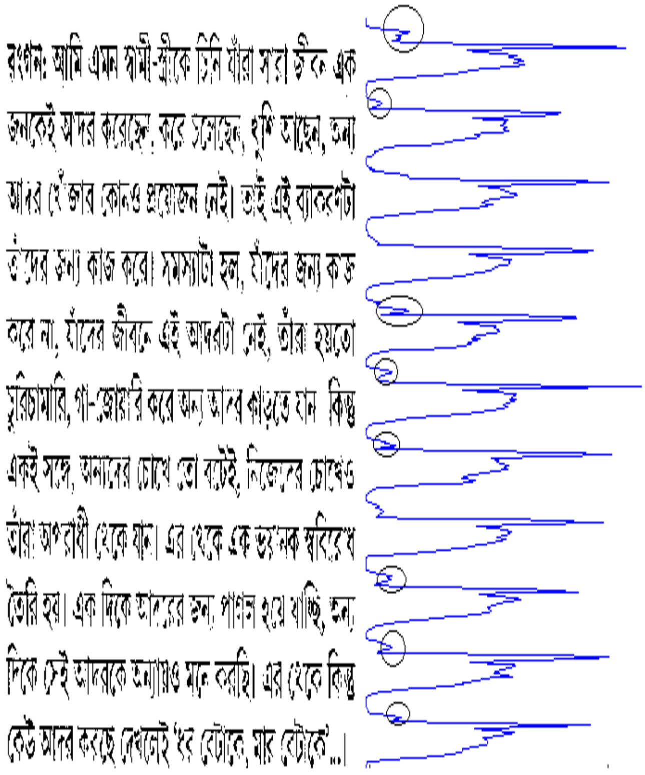

Figure 2: Horizontal projection of Bangla script document resulting under segmentation.

Algorithm 1a: Segment_Lines_Scripts(binary form of the import document)

```
begin
import text document;
apply preprocessing technique in the input document;
convert input document in to the binary matrix [] [];
extract all the strips and also calculate height of each and every strip accordingly;
if(document is from headline based script)
find the positions of all headlines;
else
find the positions of all meanlines;
end-if
$Avg[\ line\ height] = \frac{position\ of\ headlines}{meanlines}$;
for(all strips)
if(height[strip] <P1 × Avg [line height])
end-if
if(height[strip] > 40% of Avg [line height])
compute baseline [position], height[middle zone] from headline and baseline positions;
add (P2× height[middle zone]) to baseline row for finding actual line boundary ;
if(strip boundary> actual line boundary)
height[strip]- -;
```

```
loop for the same strip;
end-if
if(strip[boundary] < actual line boundary)
repeat
consider next strip;
until(strip[boundary] < actual line boundary)
end-if
end-if
end-for
end-algorithm
```

Algorithm 1b (Segment_Lines_Scripts): segmentation of uniform sized text lines of printed Indian scripts

BEGIN

Step 1: Using the *horizontal projections*, different strips denoted by $ST_1, ST_2, ST_3, \dots, ST_m$in input binary document are identified. For that, whenever $HP(i) = 0\, for\, i = 1, 2, 3, \dots, L$, it is marked as the boundary of strip line and called first row of strip $p$as $FR(Sp)$, last row of strip $p$as$LR(Sp)$. Height of the strip is calculated as $H(Sp) = LR(Sp) - FR(Sp) + 1, for\ p = 1, 2, 3, \dots, m$.

Step 2: If input document is from any headline based script, go to step 3 else go to step 4.

Step 3: Identify the position of headlines using horizontal projections. Denote the ending position of the headlines as $H_1, H_2, H_3, \dots, H_n$. Also denote the lines to be identified as $L_1, L_2, L_3, \dots, L_n$.

//number of headlines are same as number of actual lines Go to step 5.

Step 4: Identify the position of meanlines using first order differences of horizontal projection. Indicate the position of the meanlines as $H_1, H_2, H_3, \dots, H_n$.

//number of meanlines are same as number of actual lines.

Step 5: Define

$$AVG\_LINE\_HEIGHT = \frac{1}{n-1}\sum_{i=2}^{n}(H_i - H_{i-1})$$

Step 6: Set *LINE_NO* = 1 and first row of line *LINE_NO* as first row of first strip, *i.e.*,

*FR(LLINE_NO)= FR(S*1*).*

Step 7: For *i*= 1 to *m*, perform the following operations:

Step 7.1: if (*H(Si)* < (**P1** × *AVG_LINE_HEIGHT*)), *Si* is of type 1. (contains only upper zone) Repeat step 7.

//ignore current strip and go for next strip.

Step 7.2: if $(H(Si) > (0.50 \times AVG\_LINE\_\ HEIGHT))$,

Step 7.3: identify the *position of baseline*. Mark it as *BASELINE_NO*. Also set height of the middle zone as $HGT\_MID = BASELINE\_NO - HLINE\_NO$.

Step 7.4: set last row of line *LINE_NO* as *LR(LLINE_NO)* = *BASELINE_NO* + **P2** × (*HGT_MID*).

Step 7.5: if (*LR(Si) >LR(LLINE_NO)*)

set*H(Si) = H(Si) – (LR(LLINE_NO) –FR(LLINE_NO)),*

LINE_NO+ +;

set*FR(LLINE_NO) = LR(LLINE_NO*−1*)* + 1 and go to step 7.1.

Step 7.6: if $LR(Si + 1) \leq LR(LLINE\_NO)$, set *i* = *i* + 1.

Repeat step 7.6.

//for multiple lower zones.

Step 7.7: *LINE_NO++;*

Set $FR(LLINE\_NO) = LR(LLINE\_NO - 1) + 1$. Go to step 7.

Step 8: $for\ j = 1\ to\ LINE\_NO$

Display *FR(Lj)* to *LR(Lj)* as line boundaries.

END.

4. RESULT

We applied our identification scheme on 500 line samples for Devnagari and Bangla, respectively. We have selected document image from translation book, question paper, books, computer printout, multi-lingual operational, service manual, weeklies and schoolbooks containing variable font's, stiles and size. From the experiment of an the data set we noticed that 100% script segmentation in Devnagari and 99% in Bangla result of line based script segmentation, of the proposed system. Result of different script lines are shown in the following table.

Language	Data Sample	Complete Segment	Rest	%
Devnagari Script	500	500	0.00	100%
Bangla Script	500	495	5.00	99%

Horizontal Projection Profile is implemented in MATLAB 7.4. The average time taken to recognize the script image is 0.50 second on a i3 based machine running at 1.80 GHz with 01 GB RAM.

5. ACKNOWLEDGMENTS

This Paper describes a system for Line wise script segmentation of printed Devanagari and Bangla script documents. In this Paper we have presented scheme for line wise script segmentation of Indian languages such as Devnagari and Bangla. The scheme is based on Horizontal Projection Profile of printed Indian text line. The algorithm used in the paper is very simple, easy to understand and reliable for the line wise segmentation of the scripts. In the algorithm, the segmentation rate of the scripts is very fast and accurate.

6. REFERENCES

[1] U. Pal and B. Chaudhuri. Script line separation from indian multi-script documents. In International Conference on Document Analysis and Recognition, pages 406{409, 1999.

[2] U. Pal and B. Chaudhuri. Automatic identi_cation of english, chinese, arabic, devnagari and bangla script line.

In International Conference on Document Analysis and Recognition, pages 790{794, 2001.

[3] U. Pal, S. Sinha and B. B. Chaudhuri, “Multi-Script line identification from Indian documents,” Proc. of seventh Intl. conf. on document analysis and Recognition (ICDAR 2003), vol. 2, pp. 880-884, 2003.

[4] Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet, “Identification of Scripts of Indian Languages by Combining Trainable Classifiers,” ICVGIP, Bangalore, India, Dec.20-22, 2000.

[5] S Basavaraj Patil and N.V. SubbaReddy, “Neural network based system for script identification in Indian documents,” Sadhana, vol. 27, part1, pp. 83-97, February 2002.

[6] B.V. Dhandra, Mallikarjun Hangarge, Ravindra Hegadi and V.S.Malemath, “Word Level Script Identification in Bilingual Documents through Discriminating Features,” IEEE – ICSCN 2007, Chennai, India, pp.630-635, Feb. 2007.

[7] S. Chanda, U. Pal, “English, Devanagari and Urdu Text Identification,” Proc. Intl. Conf. on Document Analysis and Recognition, pp. 538-545, 2005.

[8] P. A. Vijaya, M. C. Padma, “Text line identification from a multilingual document,” Proc. of Intl. Conf. on digital image processing (ICDIP 2009) Bangkok, pp. 302-305, March 2009.

[9] Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy, “Script Identification from Indian Documents,” LNCS 3872, DAS, pp. 255-267, 2006.

[10] Zhou L, Y Lu and C L Tan, Bangla/English script Identification based on analysis of connected component Profiles, In Proc. 7th DAS, 2006

[11] S. Tsujimoto and H. Asada, 1992, “Major components of a complete text reading system”, Proceedings of the IEEE, Vol. 80(7), pp. 1133-1149, 1992.

[12] V. Bansal and R. M. K. Sinha, “Segmentation of touching and fused Devanagari characters”, Pattern Recognition, Vol. 35(4), pp. 875-893, 2002.

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

[14] B. B. Chaudhuri and U. Pal, “A complete printed Bangla OCR system”, Pattern Recognition, Vol. 31(5), pp. 531-549, 1998.

[15] 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.

[16] 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.

[17] G. S. Lehal, Optical Character Recognition of Machine Printed Gurmukhi Text, Ph.D. hesis, Punjabi University, Patiala, India, 2001.