Word Level Script Identification for a Multilingual Document

Ankit Kumar[1]
[1]Department of Computer Science & Engineering,
Ganga Institute of Technology and Management,
Kablana, Jhajjar, Haryana, India

***Abstract*— India is a multilingual multi-script country. In every state of India there are two languages one is state local language and the other is English. For example in Andhra Pradesh, a state in India, the document may contain text words in English and Telugu script. For Optical Character Recognition (OCR) of such a bilingual document, it is necessary to identify the script before feeding the text words to the OCRs of individual scripts. In this paper, we are introducing a simple and efficient technique of script identification for Kannada, English and Hindi text words of a printed document. The proposed approach is based on the horizontal projection profile for the discrimination of the three scripts. The feature extraction is done based on the horizontal projection profile of each text words. We analysed 500 different words of Kannada, English and Hindi in order to extract the discrimination features and for the development of knowledge base. We use the horizontal projection profile of each text word and based on the horizontal projection profile we extract the appropriate features. The proposed system is tested on 18 different document images containing about 400 text words of each script and a classification rate of 96.25%, 99.25% and 98.87% is achieved for Kannada, English and Hindi respectively.**

***Keywords*—Multilingual documents, Largest mean, Horizontal projection, Aspect ratio, Vertical strokes.**

I. INTRODUCTION

In Multilingual document analysis, it is important to automatically identify the scripts before feeding each text words of the document to the respective OCR system. In India, English has proven to be the binding language. Therefore, a bilingual document page may contain text words in regional language and English. So, bilingual OCR is needed to read these documents. To make a bilingual OCR successful, it is necessary to separate portions of different script regions of the bilingual document at word level and then identify the different script forms before running an individual OCR system.

In the context of Indian language document analysis, major literature is due to Pal and Choudhari. The automatic separation of text lines from multi-script documents by extracting the features from profiles, water reservoir concepts [1]. Santanu Choudhury, Gaurav Harit, Shekar Madnani and R. B. Shet has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu [2]. Chanda and Pal have proposed an automatic technique for word wise identification of Devnagari, English and Urdu scripts from a single document [3]. Word level script identification in bilingual documents through discriminating features has been developed by B V Dhandra, Mallikarjun Hangarge, Ravindra Hegadi and V.S.Malemath [4].Vijaya and Padma has developed methods for English, Hindi and Kannada script identification using discriminating features and top and bottom profile based features (English, Hindi, Kannada) [5]. B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath developed a method of Word-wise Script Identification from Bilingual Documents Based on Morphological Reconstruction(English, Hindi, kannada) [6]. Prakash K. Aithal, Rajesh G., Dinesh U. Acharya, Krishnamoorthi M. Subbareddy N. V. Has proposed a method of Text Line Script Identification for a Multilingual Document (English, Hindi, Kannada) [7].

This paper deals with word-wise script identification for Kannada, English and Hindi script pertaining documents from Karnataka, Uttar Pradesh. Script identification is done based on the features extracted from Horizontal Projection Profile of the word segment. To discriminate Kannada, English and Hindi the mean of Projection Profile Values between first and second largest and value of the point immediately after either first largest or second largest depending upon the position, which largest come earlier in the horizontal projection profile is used.

The remaining portion of the paper is organised as follow: Section 2 give some discrimination feature of English and Kannada script. Section 3 gives the pre-processing and segmentation. Section 4 covered the feature extraction. In Section 5 propose Algorithm is given. Result is discussed in Section 6.Section 7 cover the conclusion, And acknowledgement is given in section 8.

II. DISCRIMINATING FEATURES OF ENGLISH, HINDI AND KANNADA

1. In English script vertical strokes appear in the left side of the character mostly such as (B, D, H, F, R, K, P, b, h, k, l) whereas in Hindi they appear in the right side of the characters as shown in the fig 1.

Figure 1 Vertical strokes in the right side of the characters

2. In most cases height of English character is greater than its width, whereas in Kannada width of character is greater than height as shown in fig 2(a).

3. Horizontal stroke is more in Kannada script compare to the English script shown in fig 2(b).

4. Aspect ratio of English character is greater than 1, whereas in case of Kannada it is less than 1.it is the ratio of height and width.

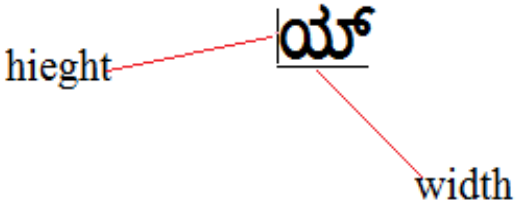

Figure 2(a) Height and width of a Kannada character

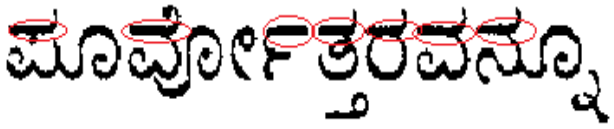

Figure 2(b) Horizontal stroke in Kannada words

5. Most of the Hindi language characters (alphabets) have a horizontal line at the upper part. In Hindi, this line is called sirorekha as shown in fig 3.

Figure 3 Headline at the upper part in the Hindi script

III. PRE-PROCESSING

The documents are scanned using HP Scanner. The scanned images are digitized images and are in gray tone. We have to convert the image into the binary image. for this we use image binarization. After binarization we perform skew detection and skew correction on the binary image.

A. Binarization

System takes input image in gray tone having pixels intensity values between (0-255) and using a thresh-holding approach converts them into two-tone images (0 and 1), black pixels having the value 1's correspond to object and white pixels having value 0's correspond to background.

B. Segmentation

White space between text lines is used to segment the text lines. The line segmentation is carried out by calculating the horizontal projection profile of the whole document. The horizontal projection profile is the histogram of number of black pixels along every row of the image. The projection profile exhibits valleys of zero height corresponding to white space between the text lines. Line segmentation is done at these points.

Similarly White space between text words is used to segment the text lines. Word segmentation is done by the vertical projection profile. The vertical projection profile is the histogram of number of black pixels along every column of the segmented line. The projection profile exhibits valleys of zero height corresponding to white space between the text words.

IV FEATURE EXTRACTION

This paper deals with word-wise script identification for Kannada, English and Hindi scripts pertaining documents from Karnataka Uttar Pradesh. Script identification is done based on the features extracted from Horizontal Projection Profile of the word segment. To discriminate Kannada, English and Hindi mean of Projection Profile Values between first and second largest and value of the point immediately after either first largest or second largest depending upon the position, the largest come earlier in the horizontal projection profile is used.

V ALGORITHM

1) For each word calculate the first (L1) and second (L2) largest value of the horizontal projection profile.

2) Calculate the Largest mean, Lm (Largest mean is the mean of projection profile between the first and second Largest including both).

(3) Find the value of the point Lp (Lp is the point immediately after the Largest (L1orL2) which come first in the horizontal projection profile).

(4) Compare the Lp with Lm:-

(a) If Lp/Lm falls in the range 0.071-0.258 then the text word is recognized as Hindi.

(b) Else If Lp/Lm falls in the range 0.258- 0.5 then the text word is recognized as Kannada.

(c)Else if Lp/Lm falls in the range 0.5-0.9 then the text word is recognized as English.

VI RESULTS

The database includes 700 text words from 30 different document images. The document images are downloaded from Google and e-news paper (Times of India, Hindustan times, Sanjevani, vijaya Karnataka, Amar Ujala) for English, Hindi and Kannada. It includes the text words both in regular and italics fonts of size varying from 9 to

14. The proposed system is tested on 21 different document images containing about 400 text words of each script and a classification rate of 96.25%, 99.25% and 98.87% is achieved for Kannada and English and Hindi respectively.

Table I show the range of Lp/Lm for Kannada, English and Hindi text words. Table II gives the identification results of Kannada and English. Table III gives the identification results of Hindi and English. Table IV gives the confusion matrix for the proposed script identification. Fig.4 shows the range of Lp/Lm for English and Kannada script. Fig.5 shows the range of Lp/Lm for English, Kannada and Hindi script.

TABLE I RANGE OF Lp/Lm FOR KANNADA ENGLISH AND HINDI TEXT WORDS

Language	Range
Kannada	0.258 to 0.50
English	0.5 to 0.96
Hindi	0.071-0.258

TABLE II IDENTIFICATION RESULTS OF KANNADA AND ENGLISH

	Kannada	English
Kannada	97.25%	2.75%
English	1.25%	98.75%

	Kannada	English	Hindi
Kannada	96.25%	2.5%	1.25%
English	0.725%	99.25%	0.025%
Hindi	0.24%	0.885%	98.875%

TABLE III IDENTIFICATION RESULTS OF HINDI AND ENGLISH

	Hindi	English
Hindi	99.65%	0.35%
English	0.33%	99.67%

TABLE IV CONFUSION MATRIX OF SCRIPT IDENTIFICATION (TRI-LINGUAL DOCUMENT)

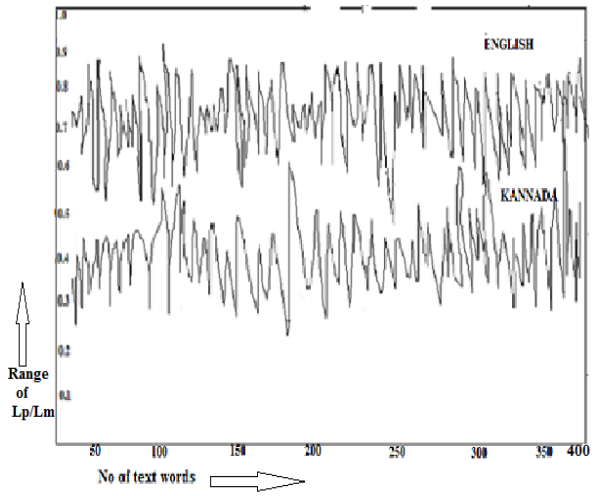

Figure 4. RANGE of Lp/Lm FOR ENGLISH AND KANNADA SCRIPT

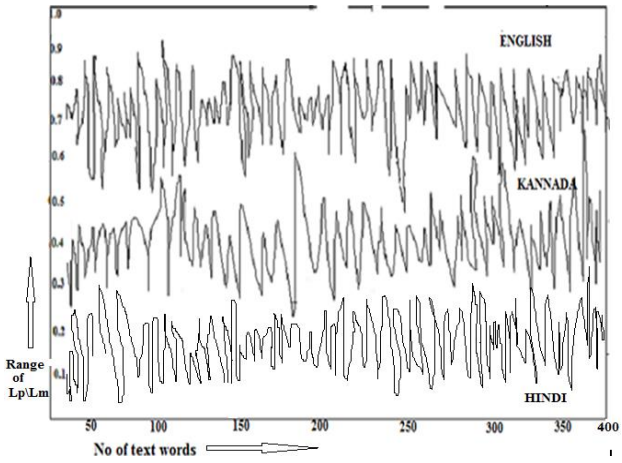

Figure 5 RANGE of Lp/Lm FOR ENGLISH AND KANNADA SCRIPT

VII CONCLUSION

In this paper, a simple and efficient algorithm for script identification of Kannada, English and Hindi text words from printed documents is proposed. The approach is based on the analysis of horizontal projection profile and does not require any character segmentation. It is based on word level segmentation. The system exhibits an overall accuracy of 98.125%. The work could be extended to character level script identification and for other Indian scripts. We can also introduce vertical projection profile based features in the approach in order to increase the accuracy of the system.

REFERENCES

[1] U. Pal, B. B. Choudhuri, "Script line separation from Indian multi-Script documents," Proc. of fifth Intl. Conf. on Document Analysis and Recognition (IEEE computer society press), pp. 406-409, 1999.

[2] 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.

[3] S. Chanda, U. Pal, "English, Devanagari and Urdu Text Identification,"*Proc. Intl. Conf. on Document Analysis and Recognition*, pp. 538-545,

[4] B.V. Dhandral, Mallikarjun Hangarge', Ravindra Hegadil and V.S. Malemathl IEEE " Word Level Script Identification in Bilingual Documents through Discriminating Features "- ICSCN 2007 *Chennai, India*, pp.630-635, Feb. 2007.

[5] P. A. Vijaya, M. C. Padma, "Text line identification from a multilingual document," Proc. *of Intl. Conf. on digital image processing (ICDIP2009) Bangkok*, pp. 302-305, March 2009.

[6] B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi V.S Malemath. " Word-wise Script Identification from Bilingual Documents Based on Morphological Reconstruction".

[7] Prakash K. Aithal, Rajesh G., Dinesh U. Acharya, Krishnamoorthi M. Subbareddy N. V. "Text Line Script Identification for Tri-lingual Document"Manipal Institute of Technology Manipal, Karnataka,INDIA

[8] M. C. Padma and P. A. Vijaya, "Identification and separation of Text words of Kannada, Telugu, Tamil, Hindi and English languages through visual discriminating features," *Proc. of Intl. conf. on Advances in Computer Vision and Information Technology(ACVIT-2007),Aurangabad, India,* pp. 1283-1291, 2007.

.