Latin Script Detection and Removal from Devanagari Document Image for OCR

Savita Pal Godara[#1], Pratap Singh Patwal [*2]
*M.Tech Scholar[#1], Associate Professor[*2]*
Institute of Engineering and Technology, Alwar, India

Abstract**—** *Document image analysis is the process or techniques used for images of documents to obtain a computer-readable description from pixel data. A document image analysis product is the Optical Character Recognition (OCR) software that recognizes text in a scanned document image. OCR makes it possible for the user to edit or search the document's contents. In this paper we proposed a novel method for identification of Latin text from Devanagari script image document. There are many documents in Devanagari where a single document page may contain English text as well with Devanagari. In bilingual documents two scripts are generally mixed together within a single text line. There are existing methods for recognition of both script but methods lack the ability to recognize multiple scripts mixed within a single text line.*

Keywords**—** OCR, Image Document, Devanagari Script, Latin Script

I. INTRODUCTION

Nowadays multimedia data specially images captured and stored increasing rapidly with the advances in information technology. In most of the document data or we can say text data are in multilingual form. In India documents data are in combination of Hindi with few of the English words or somewhere English with few of the Hindi words. For example any newspaper whether it is in Hindi or English, it has few of the second script words. Optical Character Recognition help to process captured and stored images and convert it to editable form. The purpose of OCR is to convert images into editable form which can be easily stored and used. The problem will arise when OCR is designed for one script and the document image contains few words of another script. In such conditions it is required to separate those words from the image before processing. These words can be identified separately by the specified OCR and after processing merged at the previous location. For this problem we need to identify another script word and to remove from the location. This type of pre-processing task of OCR is called script identification. In this paper we are proposed a novel method to identify Latin script from Devanagari script image document. Devanagari is used in many Indian languages like Hindi, Nepali, Marathi, Sindhi etc. More than 300 million people around the world use Devanagari script [1]. Nowadays English text is widely used as a secondary script in any Devanagari document such as newspaper, magazines, novels, office documents, school books etc. Generally English words mixed with most of the documents in Hindi, motivates us to work on this problem as shown in Figure1.

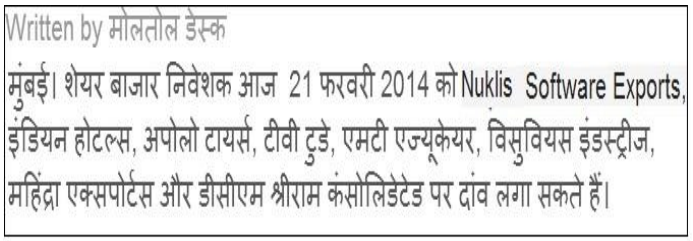

Figure 1 Bi-lingual Image Document

Script identification of Indian Scripts has been a challenging research problem in a multilingual and bilingual environment over the last few years of development of OCR. Basically the works on script identification are classified into two as local or global approach. Various works has been proposed till using local approaches [1, 2]. Generally local approaches use different features like water reservoir principle morphological features, profile, cavities, corner points, end point connectivity, top and bottom profile based features. In local approaches, the features are extracted compare and classify using different standard algorithms. The accuracy rate of the local approach depends on the pre-processing steps taken and applied in appropriate way. The very first issue occurs that features extraction perform on segmented line, segmented word and segmented character level, which are obtained only after using segmenting the underlying document image up to appropriate level. So, the success rate of identification depends on the effectiveness of the preprocessing steps such as, Line, Word and Character segmentation, noise removal etc. In this paper we proposed a world level script identification of Latin script from Devanagari script document using global features. Theses global features

are detected using projection profiles of the segmented words of both the scripts.

II. RELATED WORK

In the earlier work for Latin script identification at word level from other script document various approaches proposed. Later, Spitz presented language identification in Han-based and Latin-based scripts by using vertical position distribution of upward concavities, optical density distribution and most frequently occurring word shapes characteristics [4]. Pal and Chaudhuri worked on separation of text lines from different scripts using features like water reservoir concepts [5]. Zhou et al. presented the Bangla and English script identification by analyzing connected component profiles and head-line feature [6]. As the environment is different here for Devanagari script is mixed with it and requires a robust algorithm based on the features of their shape and direction as well.

III. PRE-PROCESSING

For the implantation testing we are input images from different sources. While image acquisition there is chances to noise in the image which need to identify and remove before the actual processing. Therefore few of the preprocessing phases are implemented such as Image Binirization Dilation & erosion, skew correction and word segmentation.

Binirization

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.

Dilation and Erosion

Dilation and erosion are two morphological operations use for the restructure the pixel elements of image. Basically Dilation adds pixels to the boundaries of elements in an image. While images capturing few of the pixel points missed and due to noise and it need to recover for appropriate recognition. In other hand erosion is the process which removes pixels on elements boundaries of the image. The number of pixels added or removed from the elements in an image depends on the size and shape of the structuring element used to process the image.

Skew Correction

Perspective disorder of scanned image due to error in acquisition process called skewed image. Skew detection and correction is the basic preprocessing step required to align the text characters in the whole image.

Word Segmentation

When the image matrix is ready to be processed, to isolate each line of the text from the whole document horizontal projection profile technique is used. A computer program scans the image horizontally to find the first and last black pixels in a line and single lines isolates as shown in the figure2. In the similar manner vertical projection profile is used for the word segmentation from a single line.

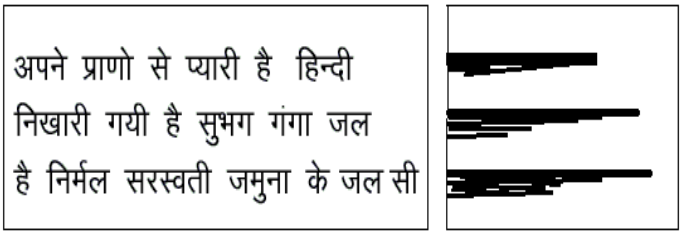

Figure 2 Projection profile for Word Segmentation

IV. PROPOSED METEDOLOGY

Proposed model is inspired from the general observation that humans are capable of distinguishing between scripts just based on simple visual prospects. Therefore different structure of word elements motivates us to find the solution to discriminate two major scripts. In general, a text structure is a complex visual pattern composed of different sub elements. Devanagari script consist a sirorekha at the top profile which is never present in the English character. In this approach we need to discriminate Latin word from Devanagari script, so we need to identify the structure of the Latin script word. As it is evident from the study that histogram are frequently used in number of techniques to identify directional features of the underlying gray scale images for their discrimination. However, in this case binarized document image has been used which contains both Devanagari as well as Latin script word. Here, we made an attempt to show that projection profile could also be used efficiently to obtain the directional features of the underlying binary image on less computational cost.

In Devanagari script all the vertical strokes present at the right most portion of the character but in other hand English character has most of the vertical strokes present at left. As another feature horizontal stroke at

the top of Devanagari word as a sirorekha but Latin word does not. In most of the Latin word there is a space between individual characters but as in Devanagari script sirorekha is there, so no space present between two characters. These are the basic feature which are be used in the algorithm using horizontal and vertical projection profile feature.

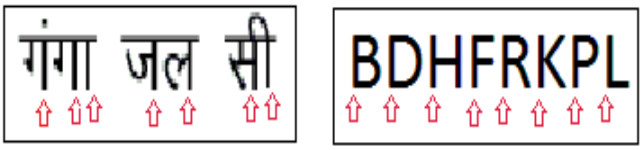

Figure 3 Devanagari and Latin Character with vertical stroke at right end and left end respectively

After segmentation of the preprocessed image following steps are taken to identify the script of the word as:

Steps for the Algorithm

1. For each word calculate the word height (H) by using horizontal projection profile.

2. Find the number of vertical strokes (VS) equal to the word height (H) by using vertical projection profile.

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

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

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

6) Compare the Lp with Lm and find number of VS:-

(a) If Lp/Lm falls in the range 0.071-0.258 and VS >1 then the text word is recognized as Hindi.

(b) Else if Lp/Lm falls in the range 0.5-0.9 and VS >1 then the text word is recognized as English.

V. EXPERIMENTAL RESULTS

In this section we evaluate the proposed method of Latin word discrimination from Devanagari using feature profiles. For implementation we used OPEN-CV 2.3 with Visual C++ using Visual Studio 2010. Over 10 image document of bilingual as well as only Devanagari script are taken from different sources such as magazines, newspapers, text books etc for the determination of range of Lp/Lm. It contains almost 400 words which are tested on the implemented model. These testing images were scanned on 300 dpi. The different experiment is designed to evaluate the efficacy of the complete system. In the step 6 of the algorithm we calculate a ratio of Lp/Lm which gives us the result matrix to identify the words. The range of Lp/Lm gives the idea to identify Latin words from Devanagari script. Final accuracy of the system tested over 20 document image which contains approx 1000 words images of both scripts.

Table 1 Scale of Lp/Lm for both scripts

S. No	Script Type	Scale of Lp/Lm
1	Latin	0.45 to 1.0
2	Devanagari	0.03 to 0.35

Table 2 Matrix for Identification of both scripts

	Latin	Latin
Latin	99.35 %	0.65 %
Devanagari	0.20 %	99.80 %

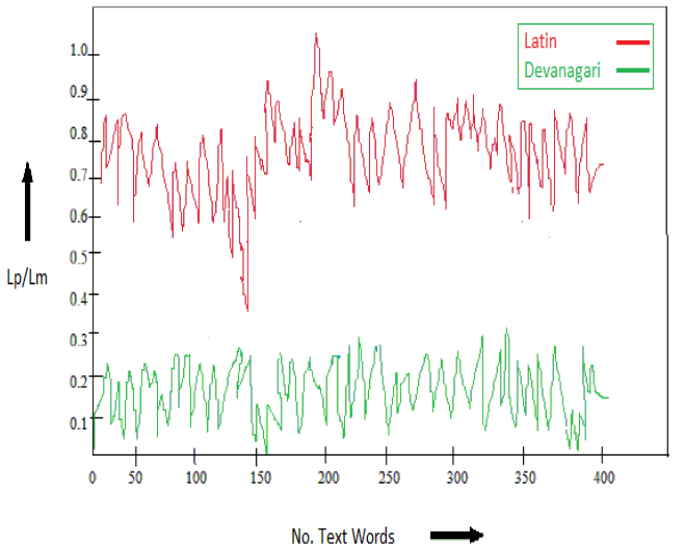

Figure 4 Graph Representing Lp/Lm scale for Latin and Devanagari Script

For the result analysis we tested our algorithm over different bilingual image with variations and designed Lp/Lm ratio scale for the identification shown in the Table 1. For the accuracy of the system different images taken as an input for the system and evaluate the performance matrix as shown in the Table2. We have generated a graph for the Lp/Lm ratio of Latin and Devanagari script shown in the Figure4.

VI. CONCLUSION

In this paper we tried to show a robust method for Latin script identification from Devanagari script document using structuring features and projections. The domains used here for implementation is the spatial domain by which system is processing. The approach is based on the analysis of horizontal and vertical projection profile and does not require any character segmentation. It is based on word level segmentation. The system exhibits an overall accuracy of more than 99%. In future, we shall study use of our algorithm for character level identification in for more than two Indian scripts.

REFERENCES

[1] Vijay Kumar , Pankaj K. Sengar , Segmentation of Printed Text in Devanagari Script and Gurmukhi Script International Journal of Computer Applications (0975 – 8887) Volume 3 – No.8, June 2010.

[2] Ankit kumar, Tushar Patnaik, Vivek Kr Verma"Discrimination of English to Other Indian Languages (Kannada and Hindi) for OCR System" AIRCC International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.2, April 2012 PP 167-175.

[3] 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. A.L. Spitz, "Determination of the script and language content of document images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 3, pp. 235–245, March 1997.

[4] U. Pal and B.B. Chaudhuri, "Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line," in ICDAR '01: Proceedings of the Sixth International Conference on Document Analysis and Recognition,Washington, DC, USA, 2001, pp. 790–794.

[5] L. Zhou, Y. Lu, and C.L. Tan, "Bangla/English script identification based on analysis of connected component profiles," in 7th IAPR Workshop on Document Analysis Systems, Nelson, New Zealand, Feb 2006, vol. 3872 of Lecture Notes in Computer Science, pp. 243–254.

[6] Benjelil, M. , REGIM-ENIS, Sfax, Tunisia Mullot, R. ; Alimi, A.M." Language and Script Identification Based on Steerable Pyramid Features" Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference Pp 716-721 , 18-20 Sept. 2012.