Identification and Removal of Devanagari Script and Extraction of Roman Words from Printed Bilingual Text Document

Ranjana S. Zinjore
Assistant Professor
KCES's Institute of Management & Research
Jalgaon

R. J. Ramteke
Professor
School of Computer Sciences
North Maharashtra University, Jalgaon

ABSTRACT
In this paper, a generalized framework has been proposed for Identification and Removal of Devangari (Marathi) Script and extraction of Roman (English) words from printed Bilingual Text document. For identification, the gray scale image is converted into binary image. After that, Sobel edge detector is applied on binary image. The morphological dilation with square structuring element is applied on image. Then labeling the connected component and with the help of visual discriminating features, Marathi words are identify. All identified Marathi words are removed from document for word level extraction of Roman Script. For Extraction of Roman words, the close neighbors, to bounding box (BB) are joined and two BB that are on the same text line in the image are group if the distance between them is less than considered threshold value. We are tested the proposed methodology on 10 different bilingual documents collected from newspapers, book text and some are manually generated. The identification accuracy obtained is 85.95%.

Keywords
Bounding Box; Morphological operation; Feature Extraction; script Identification.

1. INTRODUCTION
We Optical Character Recognition (OCR) will play an important role in transformation of paper based society to paperless electronic information society [1]. In a multi-lingual multi-script country like India has more than 18 official languages and 12 different scripts are used for these languages [2]. In India documents are written in state official language with combination of English language such as Books, Newspapers and Magazines. So, Bilingual OCR is need to read these document and to make a bilingual OCR successful, it is necessary to separate words from the bilingual document and identified the script before passes to specific OCR [3]. In this paper we used connected component approach to identified and removed Devanagari (Marathi) Script from printed bilingual document and then extracted Roman (English) words. Marathi is derived from Devanagari Script, written from left to right with no upper and lower case characters. Every character has a horizontal line at the top called as the header line [4]. English is derived from Roman script consists of 5 vowels, 21 consonants and 10 numerals. It has total 26 upper case and lower case letters [5].

2. PROBLEM IDENTIFICATION AND LITERATURE REVIEW
Optical Character Recognition (OCR) will play an important role in a society for visually impaired people when voice synthesizer is attached. A system is developed for English and other European languages having name as JAWS (Job Access with Speech) originally released in 1989 by Ted Henter, is a computer screen reader program that allows blind and visually impaired users to read the screen with text to speech output. But there is development of bilingual Optical character recognition system for printed and handwritten scripts.

K. Roy, U.pal and B. B. Chaudhuri have a lot of contribution in the development of postal automation system for Bangla and English scripts [6] and Oriya – English Scripts [7]. Lijun Zhou et al. [8] proposed a system for Bangla / English script identification using connected component profile with application to destination address block of Bangladesh envelop images. Gurpreet Lehal et al. [9] used nine structural features for script identification of Gurumukhi and Roman characters and words. Sunilkumar Sangame et. al. [10] discussed some discriminating features and voting technique as a classifier for Kannada and English script from Handwritten Bilingual document. Savita Pal Godara et. al. [11] suggested horizontal and vertical projection profile for identification of Latin script.

3. PROPOSED WORK
The architecture of the proposed work is shown in Figure 1.

3.1 Preprocessing
The experimental bilingual data have been collected from different sources as newspapers, books text, book cover and manually generated text. Color images are first converted into gray scale images and then into binary images. The small objects from the documents are removed using morphological opening followed by extracting Bounding Box (BB) form document. Finally thinning is applied on BB.

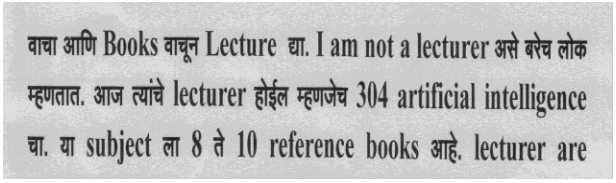

Fig 2: Bilingual Sample dataset

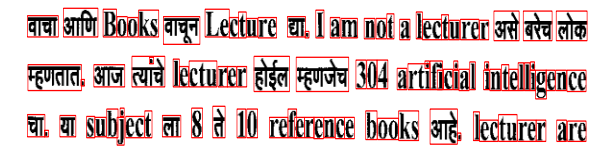

Fig 3: Bounding Box using label Connected Component

3.2 Feature Extraction
For feature extraction it is necessary to remove full stop from the word image. It is observed that full stop is present at right bottom corner of the word and the size of full stop after

thinning is not more than 6 pixels. We used two different word level structural features involves Header-line pixel count and Inter character gap.

1) Header-line Pixel Count (F1): Header-line is defined as horizontal row in upper 40% part of a word with maximum number of black pixels (maxpixel) [9]. Header line is a very important feature to distinguish Marathi words from English words based on threshold value. The threshold value (th) is considered as 43% of word length.

2) Inter Character Gap (F2): Another distinct feature of Marathi words is absence or approximately 2% of inter character gap due to words are connected with header-line/shirorekha, where as maximum English words has inter character gap. Column wise summation of white pixels (vcnt) are counted and from vcnt (final = 97*vcnt/100) we have counted only those pixels whose value is equal to zero and stored in ncnt.

3.3 Classification

For classification of Marathi words from bilingual documents, heuristic rule based approach is used and the result is shown in Table 1.

Algorithm ()

Input: Marathi and English words

Output: Marathi words

1. Compute Feature F1
2. Compute Feature F2
3. Identification of Marathi words as follows:

   if maxpixel > th OR ncnt>final

   Words = “Marathi”

   else

   Words = “English”

   End
4. Return words

The connected components (CC) which are not satisfied the above conditions, those cc are activating in blank area as shown in Figure 4.

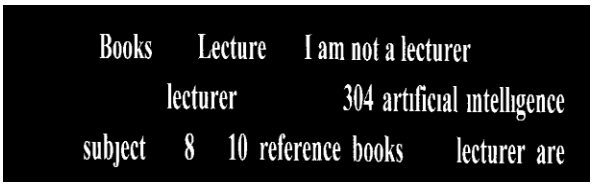

Fig 4: Separation of Devanagari words from bilingual Text

3.4 Word level Extraction

For Extracting English words form the document following algorithm is used.

3.4.1 Algorithm:

Input: Removed Devanagari text Image

Output: Extracted English words

1) The Boundary Box contain four element such as x , y (upper left corner) and xwidth, yheight .

2) Stored x and xwidth in one array name as first.

3) Stored y and yheight in another array name as second.

4) The closed neighbors to BB are joined for creating words. Two BB, that are on same text line in the image are grouped if the distance between them is less than the following threshold value th [12] (Eq. 1):

Let cnt=0;

cnt1=0;

Distbb = [];

For k = 1:length(first)

Compute: cnt = cnt+first(k,2)

cnt1 = k+1

Th = cnt / 2*n; (where n is number of BB) ---------- (1)

Distbb = [Distbb;first(cnt1,1)-(first(k,1)+first(k,2))];

End

5) Stored x values and xwidth of BB to arrays such as first1 and first2. Similarly stored value of y and yheight to another array as second1 and second2.

6) For connecting the Bounding Box; add the width and compared height (Eq. 2,3,4 & 5) of two bounding box if the distance between two bounding box are than the above mentioned threshold (Th):

Let flag = true

Finalwidth = 0

Finalheight =0

For c = 1: length (Distbb)

If Distbb (c) < Th

Xwidth1 = first2(c) and

yheight1 = second2 (c) -------- (2)

If flag == true

(Then add the First BB width + Distance between the two BB + width of next BB as follows):

xwidth2 = (first2(c) + Distbb(c)+ first2(c+1)) ------ (3)

x =c

y=c

flag = false;

Else

xwidth2 = (Finalwidth+Distbb(c)+first2(c+1)) --------- (4)

yheight2 = (second2(c+1)+(second1(c+1)-second1(y))) -(5)

End

Finalwidth = Max(xwidth1, xwidth2)

Finalheight = Max (yheight1, yheight2)

Allrect = [Allrect;first1(x) second1(y) Finalwidth Finalheight];

Else

Finalrect = [Finalrect;first1(x) second1(y) Finalwidth Finalheight];

Flag= true

End

Result is shown in Figure 5.

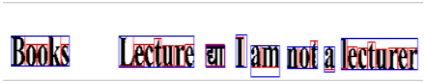

Fig. 5: closed neighbors of BB are join using blue line

4. RESULT AND DISCUSSION

The proposed algorithm has been tested on a dataset of 10 documents with varying font size with 77 lines and 474 words. For identification and removal of Devanagari script, morphological approach is used for obtaining bounding box. Full stops are removed from extracted bounding boxes and morphological thinning operation is applied for feature extraction. Header-line pixel count and Inter character gap is used as a feature for script identification and Heuristic approach is used for classification. Connected component is used to obtained pixel index list and activating the remaining boxes (English words) on blank image. For extraction of English words, the closed neighbors to BB are joined and two BB, that are on same text line in the image are grouped if the distance between them is less than the specified threshold. Table 1 shows the accuracy of Marathi word identification from printed bilingual text documents with correct classification accuracy is 85.95%.

5. CONCLUSION

In this paper a simple and efficient algorithm is used for Identification of Devanagari (Marathi) script and extraction of Roman words from printed bilingual document, which is helpful for developing bilingual optical character recognition (used in Post-office, Bank, School, Railways). In future we identified the extracted box is word or numeral.

6. REFERENCES

[1] S.Basavaraj Patil and N V Subbareddy, "Neural Network based System for Script Identification in Indian Documents", Sadhana, Special Issue on Indian Language Document Processing, Feb 2002, Vol.27, part-1, pp. 83-97.

[2] D. Ghosh, T. Dube and A. P. Shivaprasad, "Script Recognition A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 12, **(2010)** December, pp. 2142-2161.

[3] Dhandra.B.V, Mallikarjun.H, Hegadil.R and Malemathl.V.S., "*Word Level Script Identification in Bilingual Documents through Discriminating Features*", IEEE- International Conference on Signal Processing, Communications and Networking (ICSCN), Feb 22- 24, 2007, pp. 630-635.

[4] Sushama Shelk and Shaila Apte, "A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features", International Journal of Signal Processing, Image Processing and Pattern Recognition, March 2011, Vol. 4.

[5] Aarti G. Ambekar, Chhaya S. Hinge, Samidha S. Kulkarni, "Bilingual OCR for Printed English and Devnagari Text", International Journal of Research, Jan 2013, Vol. 2, Issue: 1, ISSN: 2250-1991.

[6] K. Roy, U. Pal, and B. B. Chaudhuri, "Neural Network based Word wise Handwritten Script Identification System for Indian Postal Automation", IEEE-Proceedings of International conference on Intelligent Sensing and Information Processing (ICISIP), Jan 4-7 2005, pp 240-245

[7] K. Roy and U. Pal, "Word-wise Hand-written Script Separation for Indian Postal automation", In Proc. 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR), pp. 521-526, 2006.

[8] Lijun Zhou, Yue Lu, Chew Lim Tan, "Bangla/English Script Identification based on Analysis of Connected component Profiles", In Proc. 7th IAPR workshop on Document Analysis System, New land, pp. 234-254,13-15, Feb-2006

[9] R. Dhir, C. Singh and G.S. Lehal, "A Structural Feature Based Approach for Script identification of Gurumukhi and Roman Characters and Words", Proceedings of 39th Annual National Convention of Computer Society of India (2004) December.

[10] Sunilkumar K. Sangame , R. J. Ramteke , Shivkumar Andure and Yogesh V. Gundge, "Script identification of text words from a bilingual document using voting Techniques", World Journal of Science and Technology 2012, 2(5):114-119 ISSN: 2231 – 2587

[11] Savita Pal Godara and Pratap Singh Patwal, "Latin Script Detection and Removal from Devanagari Document Image for OCR", International Journal of Computer & Organization Trends, Mar 2014, Vol.6.

[12] Lincoln Faria and Angel Sanchez, " Word- Level Segmentation in Printed Handwritten Documents",

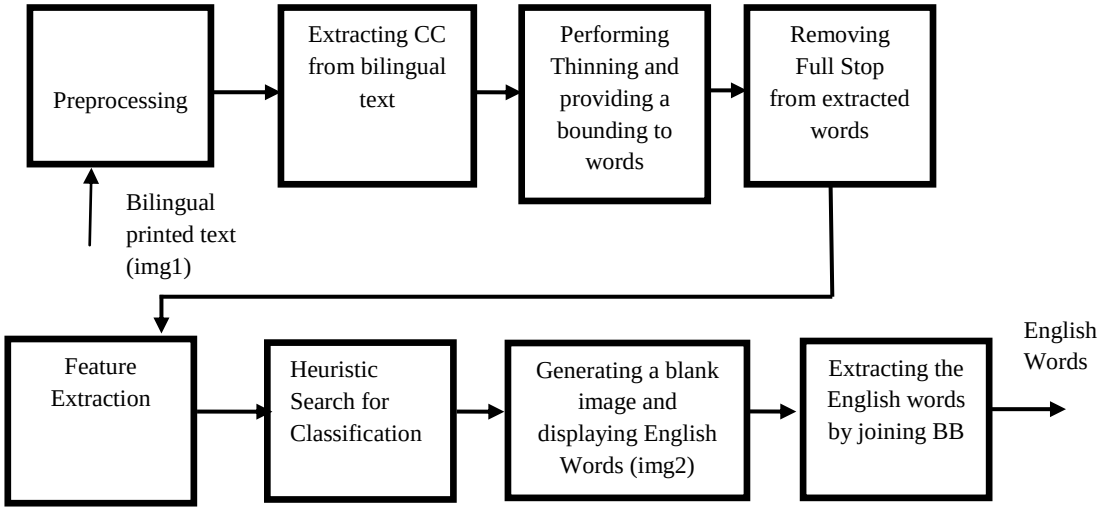

Fig. 1: Architecture of Identification and Removal of Devanagari words and Extraction of English words

Table 1: Accuracy (%) of Marathi word Identification

Bilingual Documents	**Total No. of lines**	**Total No. of words**	**Total No. of Marathi words**	**Total No. of Marathi word identified**	**Correctly Classified Marathi words**
Document1 (Manually created dataset)	08	78	45	40	**88.89%**
Document2 (Manually created dataset)	04	29	18	13	**72.22%**
Document3 (Scanned book chapter)	08	51	36	29	**80.56%**
Document4 (Scanned book chapter)	08	90	49	47	**95.92%**
Document5 (Scanned book chapter)	11	67	52	50	**96.15%**
Document6 (Newspaper Advertisement)	05	38	28	21	**75.00%**
Document7 (Newspaper text)	13	76	65	46	**70.77%**
Document8 (book cover text)	06	16	03	03	**100%**
Document9 (book cover text)	08	13	05	04	**80%**
Document10 (book cover text)	06	16	03	03	**100%**