International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 03 | June-2015 www.irjet.net p-ISSN: 2395-0072

Bilingual Script Identification of Printed Text Image

Inderpreet Kaur[1], Saurabh Mahajan[2]

[1] P.G. Scholar, Dept. of ECE, Chandigarh University, Gharuan, Punjab, India
[2] Assistant Professor, Dept. of ECE, Beant College of Engineering & Technology, Gurdaspur, Punjab, India

Abstract - *Now days script identification is desirable and challenging task in optical character recognition system for bilingual or multilingual text. Some methods have already been published in this field, but work is still in progress. For better accuracy it is vital to distinguish scripts before proceeding for classification of characters. This paper presents a work in identification of English and Punjabi scripts at line level through headline & characters density features. To classify characters in each script system accesses their individual OCRs and classifies them through unique histogram projection profiles & different number of holes features. The system is trained for English characters with Arial font and for Punjabi characters with Gurbanikalmi font. The proposed approach has been thoroughly tested for different font size images and the result presents desirable accuracy.*

Keywords: Bilingual OCR, Script identification, English, Punjabi

1. INTRODUCTION

Most of the successful effort has been made by researcher in the field of optical character recognition since decades and almost have achieved a great success to recognize monolingual printed and handwritten text images. Nowadays monolingual OCRs are failed to recognize bilingual or multilingual text. Hence expansion of OCR for recognition of bilingual or multilingual text images becomes a requisite and challenging area of research. Therefore to enhance OCR for recognition of bilingual or multilingual text, it is essential to distinguish different scripts in the text image before providing them to individual OCR systems. This addresses the need of developing script identification techniques at paragraph, line and word level [8]. Script identification technique makes convenient to analyze and recognize a text by choosing suitable modalities of OCR. Some methods have already been presented for bilingual or multilingual text images dealing with European and oriental scripts [5], English, Hindi and Kannada [9] etc. As such proposed work describes identification of English and Punjabi scripts at line level and access their specific OCRs for classification of characters each script.

Ambekar *et al* developed a method to separate and identify English and Devnagari scripts by using zoning and histogram feature and classify the characters using K-NN classifier [1].Jindal and Hemrajani described a method for script identification of printed images at line level using DCT, PCA and they tested this method for 11 major Indian languages [6].Prakash *et al* proposed an approach for script identification of Hindi, Bangla, Telgu and Kannada [7]. Gupta *et al* described some special features of Kannada, English and Hindi Characters for script identification [12].Mohanty and Bebartta explained a comparative analysis of classifier accuracies for English and Oriya printed documents [10].Dhanya *et al* used spatial spread of pixels and Gabor filter based technique for script identification of Tamil and English scripts [3].

The whole above reported piece of work accomplish a study of script identification for printed text images. This study was helpful to understand and to design script identification method for English and Punjabi scripts at line level. The headline feature together with character density is used to identify script as shown in Fig. 1.

ਦਰਖਤ ਬਸ
HAVE FUN
Have Fun

Fig -1: Illustration of headline and character density features of scripts

As it is clear from figure that numbers of Punjabi characters present in a line are less than that of English characters and headline is absent in English characters. The detail explanation of script identification is laid down in section III.

Section II involves study of different scripts to learn characteristics of scripts. The description of proposed work which focuses on script identification is presented in Section III. Section IV covers discussion on results which are obtained to show accuracy efficiency of proposed method. Section V involves conclusion of proposed method and future scope.

2. STUDY OF DIFFERENT SCRIPTS

The Punjabi is most popular language of Punjab and English is popular language of almost whole world. The Punjabi script is basically an advanced form of Gurumukhi script which was popularized by Guru Angad Dev ji. The writing style of both (English & Punjabi) scripts is from left to right. The Punjabi script contain 35 characters as shown in Fig. 1 where as English language contain 26 characters as shown in Fig. 2. The English script contains both uppercase and lowercase characters but this concept is not valid for Punjabi script. Punjabi word has headline at top of characters which is absent in English. The explanation of features which form basis of recognition of characters in respective OCRs is laid down in section III under feature extraction heading.

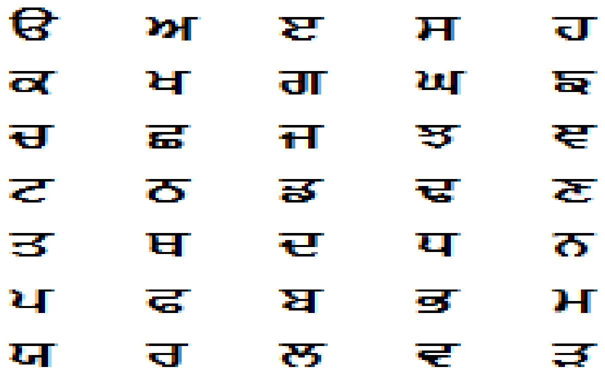


Fig -2: Punjabi characters

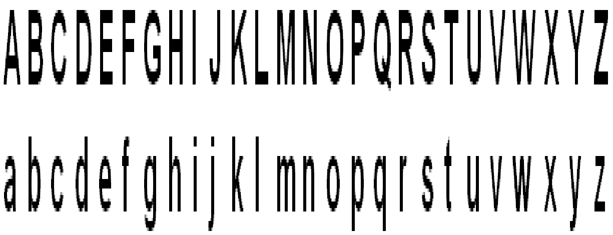


Fig -3: English uppercase and lowercase characters

3. PROPOSED WORK

To identify the bilingual input text image the block diagram of proposed script & character identification system is shown in Fig. 4.

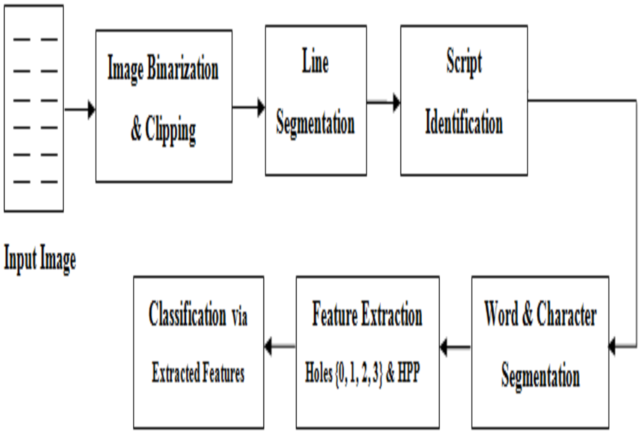

Input Image
Image Binarization & Clipping
Line Segmentation
Script Identification
Word & Character Segmentation
Feature Extraction Holes {0, 1, 2, 3} & HPP
Classification via Extracted Features

Fig -4: Block diagram of proposed script & character identification system

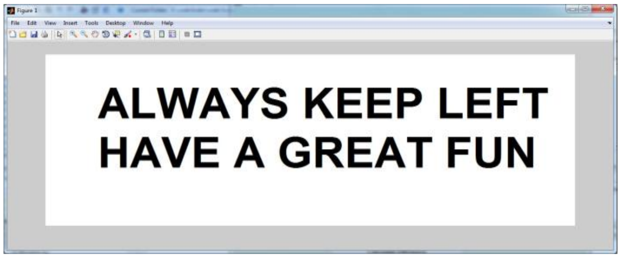
3.1 Image Binarization and Clipping

The input of an OCR may be a RGB or grayscale image so it needs to binarize the image because to identify individual glyphs and to reduce memory space. Binarization process works with thresholding to get a binary (black & white) image [2]. The pixel value above threshold is classify as white and below threshold is classify as black.

After that clip the binary image which is used to remove the sparse background pixels around the text foreground pixels. The next step is line segmentation of text image.

3.2 Line Segmentation

Generally segmentation process can be performed in three steps as lines, words and characters. The demonstration for line segmentation is shown in Fig. 5.



ALWAYS KEEP LEFT
HAVE A GREAT FUN

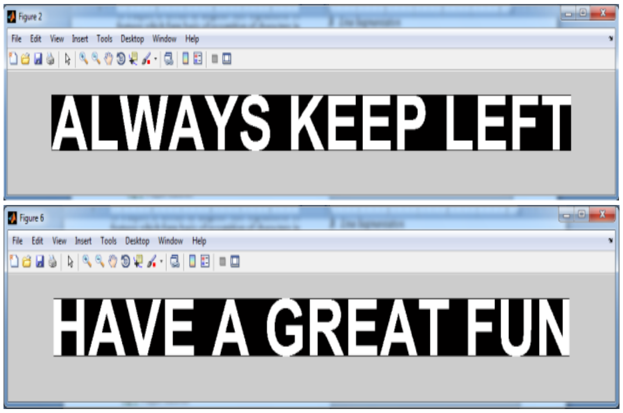


Figure 2
File Edit View Insert Tools Desktop Window Help
Figure 6
File Edit View Insert Tools Desktop Window Help

ALWAYS KEEP LEFT

HAVE A GREAT FUN

Fig -5: Demonstration for segmented lines

Line segmentation of both scripts is achieved by taking horizontal projection profile of text image [11]. The gap between two rows is detected by measurement of top to bottom row-wise sum of pixels. If sum of pixels in a row is found zero that is consider as gap between two rows and thus image is segmented from minimum row to last processed row. Repeat the process till all rows of image are segmented and image is ready for script identification.

3.3 Script Identification

Script identification is essential process to distinguish between two or more scripts which are printed in bilingual text image. Also provides knowledge for selection of particular OCRs to classify characters in each script individually. Identification of scripts can be achieved on the basis of features of scripts which are efficient to differentiate the scripts.

Here in this approach headline (presents in Punjabi and absent in English) and character density (more in English line) features are used for scripts identification. The vertical projection of each line is analyzed to locate a headline at top of characters and horizontal projection of line is analyzed to examine character density in a line as shown in Fig. 6.

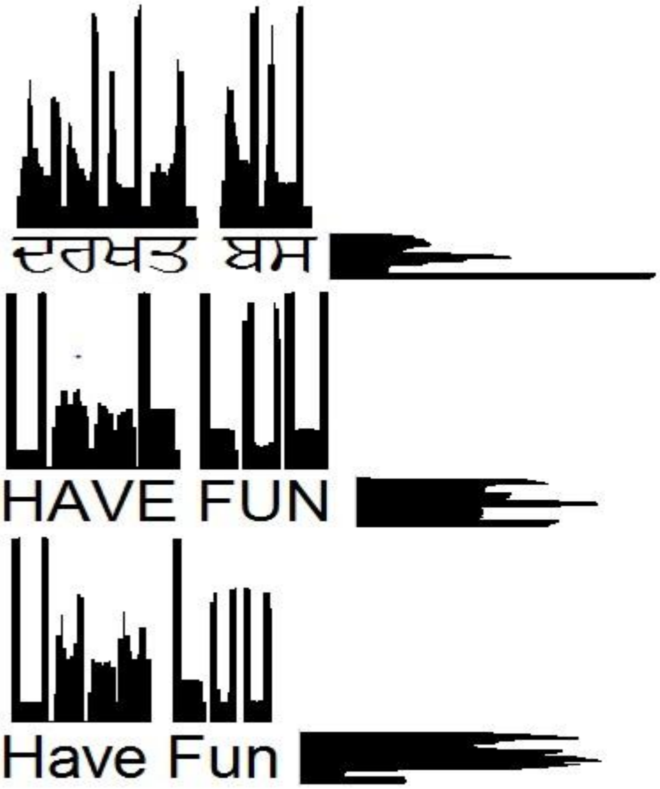


Fig -6: Horizontal & vertical projections of lines (a) Punjabi line (b) English uppercase characters line (c) English lowercase characters line

An observation of above figure reveals that each line has unique vertical and horizontal projection profile. The vertical projection of Punjabi line shows more pixels concentration while in English there are narrow gaps due to absence of headline. Similarly horizontal projection shows that English character density is more than Punjabi, because number of English characters presents in a line is comparatively more than that of the Punjabi characters. These features allow the English and Punjabi characters to be handled independently from each other. After successful script identification call individual OCRs for characters classification which first perform word & character segmentation as explained in next heading.

3.4 Word & Character Segmentation

To segment words in Punjabi line a vertical projection profile is calculated. The empty space between two words is analyzed by taking sum of pixels column-wise. As sum found zero line image is segmented from minimum row to last processed row and we got a segmented word [11]. Now process remaining words by using same algorithm till all words are segmented from a line. Due to absence of headline at top of English characters in a word the space between character and words (more width) create confusion in words segmentation. Hence in English word segmentation before proceed for vertical projection profile the first step is to define a space threshold value to overcome confusion. After this measure a vertical projection to locate a space which exceeded threshold value and separate the word from a text line. Repeat the whole process to segment the all words in a line. The word segmentation is shown as below in Fig. 7.

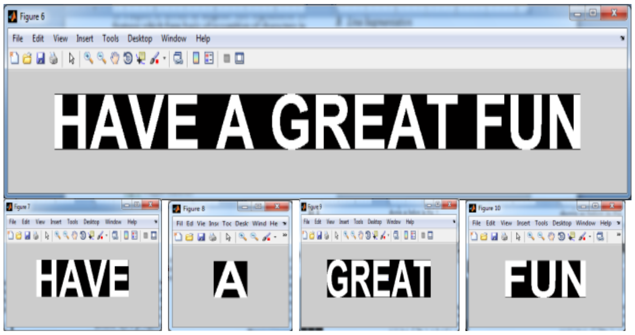


Fig -7: Demonstration for segmented words

The preceding level of segmentation is to segment the characters from each segmented word. English word consist a gap between characters. The vertical projection profile is analyzed to segment characters as same above mentioned way. The segmented characters are shown in Fig. 8. The Punjabi characters are connected by means of headline in each word. To locate character pixels in word the width of headline is consider as threshold value. Measure the vertical projection profile of word image to detect columns having value greater than threshold value and then segment columns from word image.

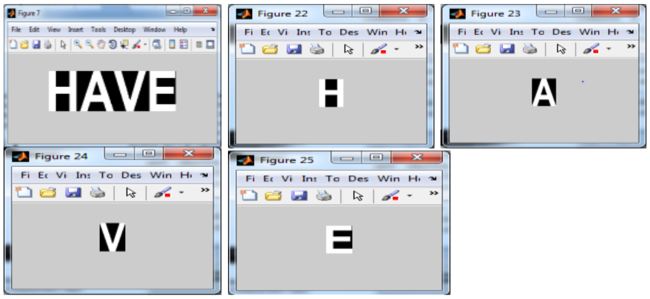


Fig -8: Demonstration for segmented characters

3.5 Feature Extraction

Next step is to extract features from single segmented character so it further passed to feature extraction level. Features disclose important attributes of character and left out unimportant attributes. The feature extraction procedure makes memo of intangible features present in a character such as structural features and statistical features [4]. The structural features which are usually geometric (end points, joints, number of cavities etc) and statistical features usually are topological (histogram projections, connectivity etc) [4]. In presented work histogram projection profiles (HPP) and number of holes features has been extracted. Histogram projection count sum of foreground pixels row wise and column wise and give a unique projection profile of each character due to different structure of each character. The number of holes features is used to group training database into four groups as 0, 1, 2, 3 holes. Hence while recognition, the test character can be retrieve from a corresponding group instead of whole database. The main aim of number of holes feature is to reduce computation power.

3.6 Classification

Classification of characters is essential step of optical character recognition system. Till now a lot of classification methods have been reported by researchers, but implementation of classifier depend on application requirements. Some classifier classifies characters by clustering method and other by feature extraction methods etc. In our work, classification of characters has been achieved via extracted feature. The system matches the features of input character which are extracted in previous stage with stored features of character in corresponding groups and give an accurately classified character. All the results saved in a text file.

4. RESULTS

For training of OCR data sets are organized for English upper & lowercase alphabet with Arial font and font size 10 to 72 as shown in Fig. 9 and Fig. 10. Alike for training of Punjabi OCR data sets are arranged for 35 Punjabi characters with Gurbanikalmi font and font size 12 to 72 as shown in Fig. 11.

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z

Fig -9: Training directory used for English uppercase alphabets

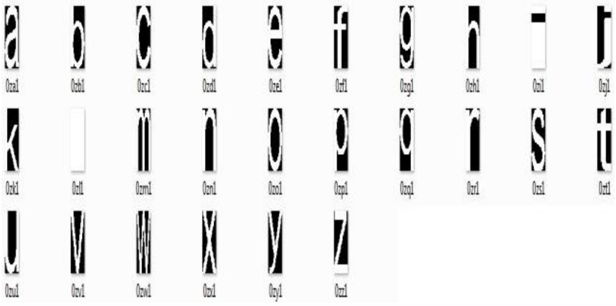


Fig -10: Training directory used for English lowercase alphabets

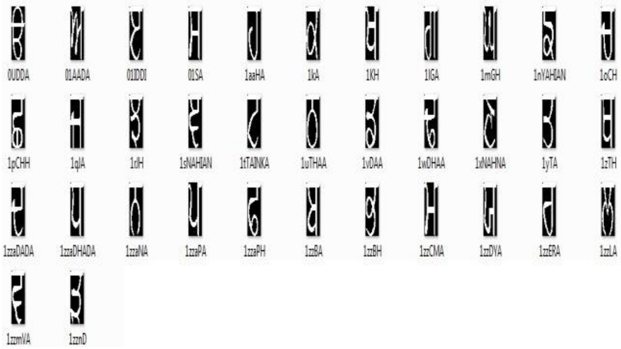


Fig -11: Training directory used for Punjabi characters

A collection for training and test sets of 1200 samples has been considered. These samples are in the form of grayscale images. Different font size images have been considered for experiments results in our proposed work. Our bilingual OCR system provides results after processing test samples through binarization & clipping, line segmentation, script identification, word & character segmentation, feature extraction and classification modules. In table 1 the results of script identification and character classification in the form of percentage are given. Results of script identification show an attractive accuracy for both Punjabi & English scripts using headline and character density features. The results of character classification have been analyzed after being sent different scripts to individual OCRs. Table 1 third column shows encouraging results for character classification with the effect of different font size images.

Table-1: Accuracy Results for Script Identification and Character Classification in %age

Consider Scripts	Script Identification (in %)	Character Classification (in %)
English	91.8	87
Punjabi	89.7	85

5. CONCLUSION

The process for script identification has been successfully implemented in our proposed work. The separation of English & Punjabi scripts through vertical projection profile is done by detecting headline and through horizontal projection profile is done by detecting number of characters present in a line. The identification of scripts is desirable because training with combined (i,e. Punjabi script and English script) database decreases accuracy and also increases recognition processing time. Here, an effort made for identification of Punjabi and English script with dissimilar training set. The efficient script identification is very important and an attempt is made to accomplish by proposing following module; preprocessing, segmentation script identification, feature extraction and classification process. The proposed system used for experimentation of printed text with different font size images. The results presented in table.1 shows achievable accuracy of proposed method. In future, other language scripts can employ for identification. One can enhance their system for handwritten text also.

REFERENCES

[1] A. G. Ambekar, C. S. Hinge and S. S. Kulkarni, "Bilingual OCR for Printed English and Devnagari Text," *Indian Journal of Research*, vol. 2, Jan 2013.

[2] A. Sharma and D. R. Chaudary, "Character Recognition Using Neural Network," *International Journal of Engineering Trends and technology*, vol. 4,April 2013.

[3] D. Dhanya, A. G. Ramakrishan and P. B. Pati, "*Script identification in printed bilingual documents*," *Springer*, Feb. 2002, vol. 27, pp. 73-82.

[4] G. S. Lehal and C. Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script," *Vivek Bombay*,1999.

[5] J. Ding, L. Larn and C. Y. Suen, "Classification of Oriental and European scripts by Using Characteristics Features," 12th *International Conference on Document Analysis and Recognition*, pp. 1023,2013.

[6] M. Jindal and N. Hemrajani, "Script Identification for printed document images at text-line level using

DCT and PCA," *IOSR Journal Computer Engineering*, vol. 12, pp. 97-102, Aug. 2013.

[7] O. Prakash, V. Shrivastava and A. Kumar, "An Efficient Approach for Script Identification," *International Journal of Computer Trends and Technology*, vol. 4, June 2013.

[8] R. Rani and R. Dhir, "A Survey: Recognition of Scripts in Bi-Lingual/Multi-Lingual Indian Documents," *National Journal of PIMT Journal of Research*, vol. 2, no. 1, pp. 55-60, March-Aug. 2009.

[9] S. B. Patil and N. V. Subbareddy, "Neural network based system for script identification in Indian documents," *Springer*, vol. 27,pp. 83-97, Feb. 2002.

[10] S. Mohanthy and H. N. D. Bebartta, "A Comparative Analysis of Classifiers Accuracies for Bilingual Printed Documents (Oriya-English)," *International Journal of Computer Science and Information Technologies*, vol. 2, pp. 916-923, 2011.

[11] S. Taha, Y. Babiker and M. Abbas, "Optical Character Recognition of Arabic Printed Text," *IEEE Conference on Research and Development*, pp. 235-240, Dec. 2012.

[12] V. Gupta, G. N. Rathna and K. R. Ramakrishan, "A Novel Approach to Automatic Identification of Kannada, English and Hindi Words from a Trilingual Document," *International Conference on Signal and System*, 2006.