International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 6–June 2013

An Efficient Approach for Script Identification

Om Prakash[1] , Mr.Vineet Shrivastava[2], Dr. Ashish Kumar[3]

Department of Computer Science and Engineering
[1]M.Tech Scholar, IFTM University,Moradabad (U.P.), India

Department of Computer Science and Engineering
[2]A.P., BBDIT GHAZIABAD (U.P.), India

Department of Computer Science and Engineering
[3]Prof, IFTM University, Moradabad (U.P.), India

***Abstract*— There are a large number of different approaches to recognize the scripts currently available in OCR System. In this report we look to identify the script of multi-languages. In the proposed script identification system, we have considered four Indian languages such as Hindi (Devanagari), Bangla, Telugu, Kannada. This system will let document images to accurate scan with higher accuracy. In this context, we modeled script identification of multilingual document using horizontal projection profile based analysis with head line features. A database of 450 text words of Hindi, 450 text words of Bangla, 450 text words of Telugu and 450 text words of Kannada are used for experimentation. The proposed system yields the 97.83 accuracy with four specified languages. Since script identification plays an important role in analyzing the printed documents.**

***Keywords*— OCR, Multi-script recognition, Binarization, Line Segmentation, Horizontal projection profile.**

I. INTRODUCTION

Nowadays, we are living in the age of computer era. So we need to store paper documents in the form of the electronic documents to storage and facilitate easy communication. Generally, the usage of paper documents is still required in most of the communications. Basically, to the communication in the world wide we used the fax machine. Also, the important point is that, normal paper is a very comfortable and easy medium to do the communication.

So, we need to have software, those can extract automatically and maintain data and information from paper documents. That information can be used for later retrieving those documents from the stored database. The approaches to solve these kinds of tasks are integrated under the general heading of document image analysis, which has been a fast growing area of research in recent years. [4]

The major operation of document image analysis is automatic extraction of text information from the paper document image. That can be achieved by a software tool that is Optical Character Recognition (OCR), which can be defined as the mechanism of manipulating the optically scanned text by the system. So we need an OCR system that can identify multi script because in multi-lingual country like India that covers 18 regional languages derived from 12 different scripts.

II. ANALYSIS OF PROBLEM

All previously literatures focused on the development of new approaches or on the improvement of existing approaches which work for some specific product. As a Result, a generalized approach to the problem has not been considered which simulates all classes of problem under a common scenario. We have focus on the some essential factors which need to consider before designing a multi-script identification approach for any multi-lingual product. These factors are Complexity in pre-processing, Complexity in feature extraction and classification, Computational speed of entire scheme, Sensitivity of the scheme to the variation in text in document (font style, font size and document skew), Performance of the scheme, and Range of applications in which the scheme could be used [6].

The problem is in identification of a Multi-script document see figure 1.

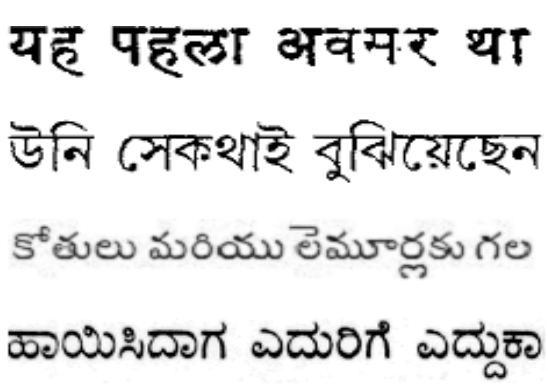

Fig. 1: Multiscript document

Performance of the approach consists the accuracy reported and selection of testing data. Previously script classification approaches can be classified into two broad categories: local and global approaches. By local approach can be defined as which analyses a scanned document image as a list of connected components like line, word and character and these components require segmentation of the image as a preprocessing step. By global approach consider an approach which handles analysis of regions comprising at least two lines. It not considers the fine segmentation.

III. PROPOSED WORK

In previous Optical character recognition system, if we provide a Multi-script document it would not identify that script with more accuracy and provides a degraded output in more times. In this research our main focus on the implementation of an approach that can help the Optical character recognition system in multi--script identification of a scanned document. In this mechanism, we will perform on the four multi script document as an input such as Devanagari, Bangla, Telugu and kannada script lines.

Input Scripts:

1. Devanagari.

2. Bangla

3. Telugu

4. Kannada.

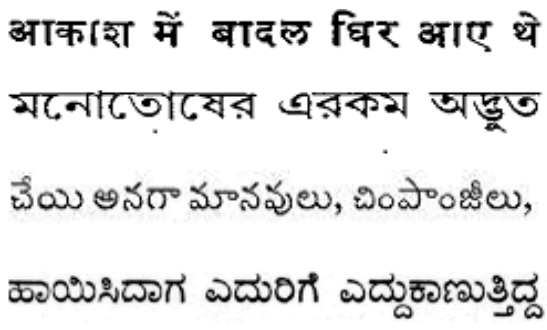

Fig. 2: Sample Input image

Proposed Framework: To identify the multi-script input sample document, we have used this general framework given below.

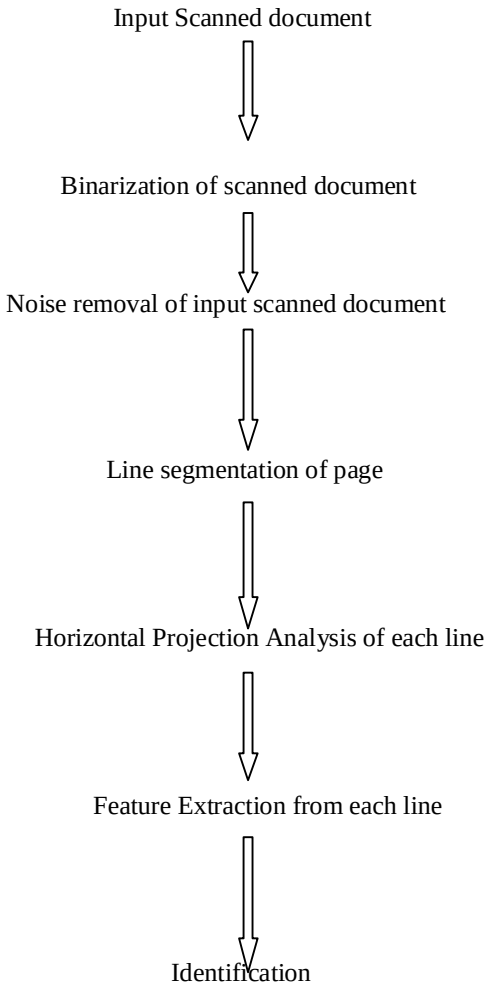

Fig. 3: Proposed approach framework

3.1: Sample Multi-Script Input Image: - In our analysis we have taken four scripts. The input scanned document Image consists of four different script languages respectively such as Devanagari, Bangla, Telugu and kannada script.

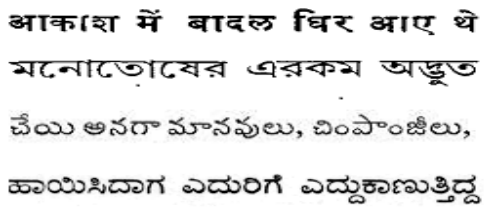

Fig. 4

3.2:Binarization: - This step shows the scanned images in (ON, 0) black and (OFF,1) white form. In the given snap shot we can see that the background is in black and symbols in white. While in normal scanned document images symbols are in black and background is in white. We have performed this conversion to pass the scanned image as an input for Horizontal projection profile.

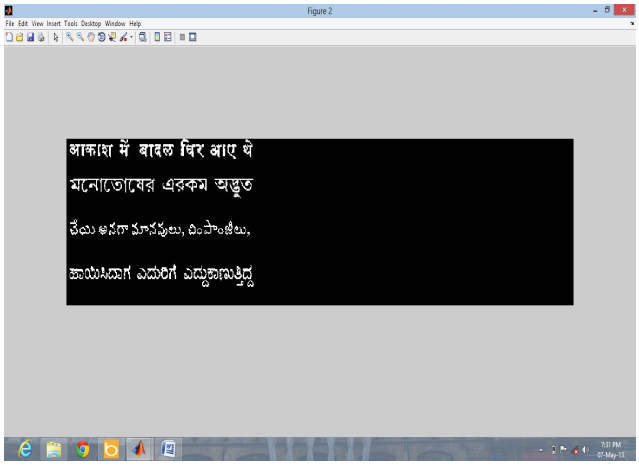

Fig. 5

3.3: Noise removal:- Noise can be defined as the dots and other unwanted marks on a scanned image document. This noise removal process is more important because our method is based on plotting peak values and gap count on a page and we will find out ratio of (ON) black and (OFF) white pixels. so we need this removal process in our framework because we have to used exact ratio of pixels.

i)Sample Multi-Script Input Image with Noise:

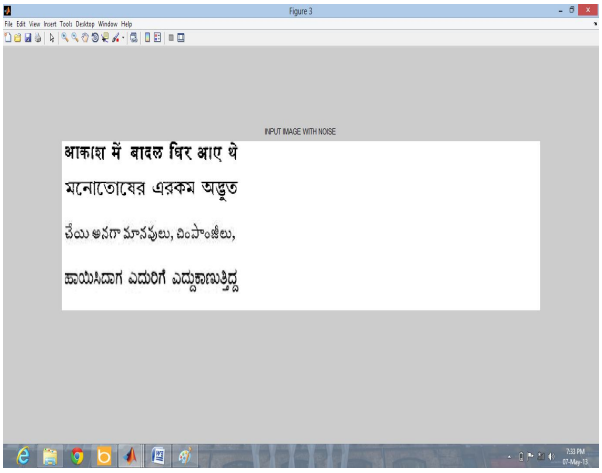

Fig. 6

ii)Sample Multi-Script Input Image without Noise:

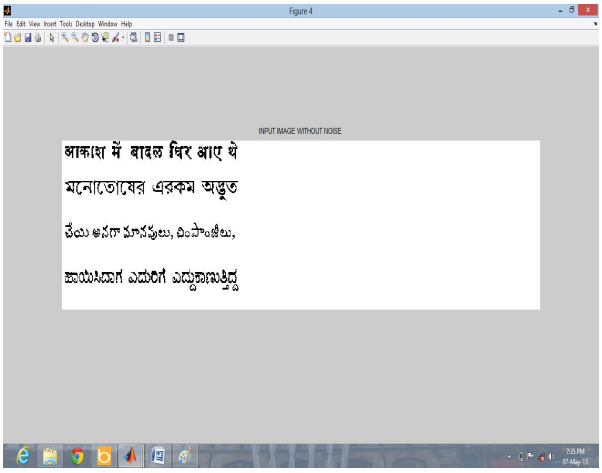

Fig. 7

3.4:Plotting of horizontal projection analysis:- In this phase, we plot the outputs of horizontal projection profile that represent the scanned image document data in histogram form and that represents the characters in the black color and background in white color and those colors shows the highest and lowest peak.

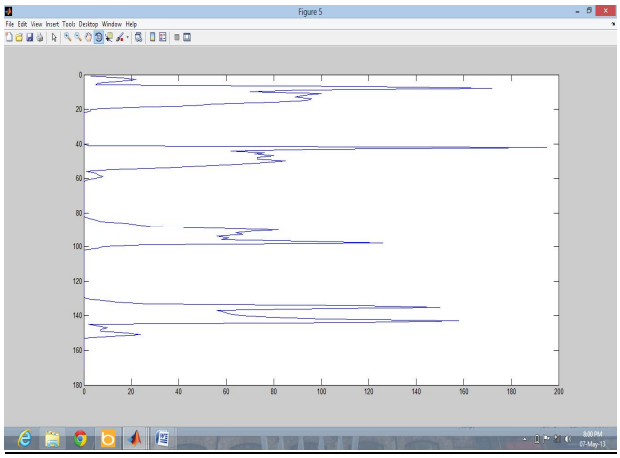

Fig. 8

3.5: Line Segmentation:- In line segmentation, white space between text lines is used to segment the text lines. It is carried out by calculating the horizontal projection profile of the scanned document. This approach is used to represent the images in the form of histogram. It represents the number of ON (black) pixels along every row of the image.

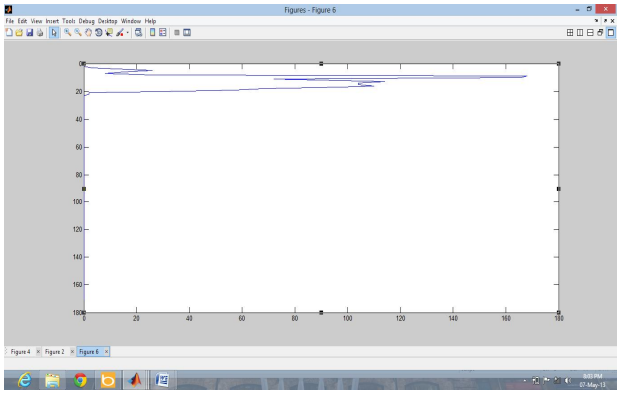

Fig. 9 Devanagari Line Segmentation

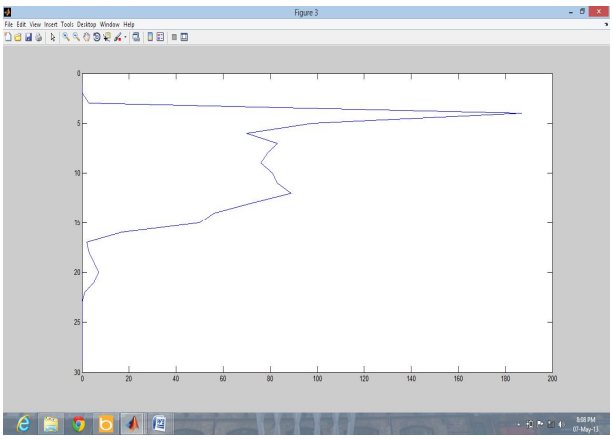

Fig. 10 Bangla Line Segmentation

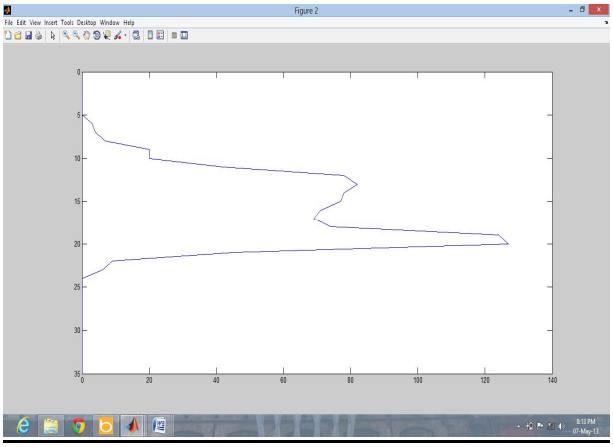

Fig. 11 Telugu Line Segmentation

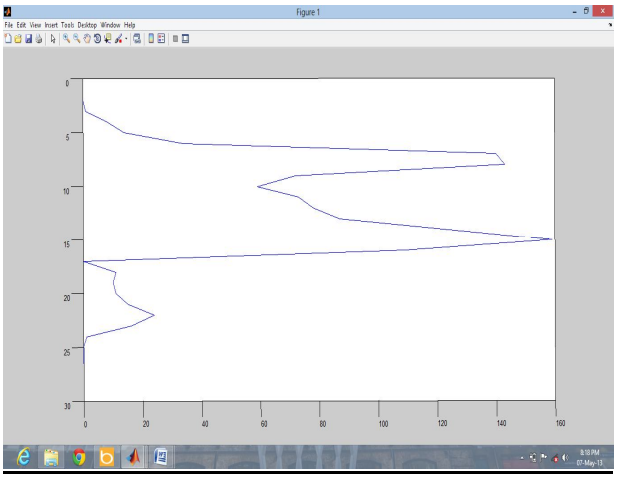

Fig. 12 Kannada Line Segmentation

3.6: Horizontal projection analysis:

a) Feature extracted from the scanned document image is 1st maxima and the 2nd maxima.

b) Now, find out the 1st maxima and the 2nd maxima of the scanned segmented images.

c) After that we have to compare the results of 1st and 2nd maxima and produce a knowledge base classifier (Rule).

d) Now, Calculate the valley mean of 1st maximum and 2nd maximum i.e. Vm.

e) Find the value of the point next to 1st maxima. i.e. Vp

f) Compare the ranges of Valley mean (Vm) and value of the point(Vp).

g) Now find out the value of Range = Value of the point(Vp)/Valley mean(Vm)

Table 3.6.1: Devanagari

Vp	Vm	Range
165	166	.9939

Table 3.6.2: Bangla

Vp	Vm	Range
89	138	.6449

Table 3.6.3: Telugu

Vp	Vm	Range
81	103	.7864

Table 3.6.4:Kannada

Vp	Vm	Range
140	150	.9333

3.7 Results :

Our proposed method implemented in MATLAB. Finally we can discriminate the languages on the basis of each script range value.

1. For Devanagari script the value of range lies between 0.96 and 1.14.

2. For Bangla script if the value of range lies between 0.01 to 0.68.

3. For Telugu script the value of range lies between 0.68 and 0.84.

4. For Kannada script the value of range lies between 0.84 and 0.96.

Table 3.7 Experimental calculation of accuracy

Script Lines	No. of Script Lines	Correctly Identified	Accuracy Percentage
Devanagari	450	448	99.55
Bangla	450	442	98.22
Telugu	450	438	97.33
Kannada	450	433	96.22
Overall	1800	1761	97.83

IV. CONCLUSION

In the overall work, we have applied the horizontal projection profile method to identify the multi-script languages. That provides the more efficient and accurate result for Script Identification with higher accuracy. This proposed system is performed on twenty different scanned document images that containing about 450 text lines of each specified script and in this process overall classification rate is achieved approximately 97.83%.

Table 6.1.1 Comparison with previous approaches

Methodology	Dataset	Accuracy
Water reservoir apporach	400	96.72%
Global approaches	300	97.27%
Proposed Approach	450	97.83%

V.REFERENCES

[1] Priyanka P. Yeotikar, P.R. Deshmukh, "**Script Identification of Text Words from Indian Document through Discriminating Features"** International Journal of Computer Applications, (0975 – 8887), 2013.

[2] Huanfeng Ma and David Doermann, "**Word Level Script Identification for Scanned Document Images",** PP. 135-140, 2012.

[3] 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, 2(5):114-119, ISSN: 2231 – 2587, 2012.

[4] M Swamy Das, D Sandhya Rani, C R K Reddy, A Govardhan, **"Script identification from Multilingual Telugu, Hindi and English Text Documents",** International Journal of Wisdom Based Computing, Vol. 1 (3), 2011.

[5] M. C Padma, P. A Vijaya, **"Script Identification from Trilingual Documents Using Profile Based Features",** International Journal of Computer Science and Applications,
Technomathematics Research Foundation, Vol. 7 No. 4, pp. 16 - 33, 2010.

[6] Gopal Datt Joshi, Saurabh garg, and Jayanti Saraswat,"**Script Identification of Indian Documents"**, LNCS 3872, PP.255-267, 2006.

[7] U.Pal, **"Automatic Script Identification: A survey**", VOL 16, PP-26-35, 2006.

[8] U.Pal, S.Sinha and B.B chaudhary**," Multiscript Line Identification from Indian documents**, Published in seventh international conference on Document Analysis and Recognition, ICDAR, 2003.

[9] U.pal and B.B chaudhary,"**Automatic Separation of different script Documents**", Published in Indian Conference on Computer-vision, Graphics and Image processing, PP 141-146, 1998.

[10] Santanu Chaudhary,Rabindra Seth ,"**Trainable Script Identification Strategies Of Indian Languages**", Published in fifth International Conference on Document Analysis and Recognition, 1999.

[11] Aspitz**," Determination of the script and language context of document images**", Published in IEEE Trans PAMI, VOL 19 NO3, PP 235-245, 1997.

[12] U.pal and B.B chaudhary**,"Automatic Separation of Roman, Devanagri and Telugu Script lines**", Published in advances in pattern Recognition and digital techniques, PP 447-451, 1999.

[13] Arvind Kumar Patel, Ashok Kumar Dubay and Vineet Shrivastava, **"Developing an optimized solution for script identification processes in a multilingual document using OCR"**, ISSN-2278-6643, 2012.

[14] www.mathworks.com

[15] www.matlabcentral.com