Script Identification for Tri-Lingual Image Document

Anil Kumar Dahiya
(SM-IEEE,ACM)
Dept. of CSE
Manipal University Jaipur
Rajasthan, India

Vivek kumar Verma
Dept. of CSE
Manipal University Jaipur
Rajasthan, India

ABSTRACT
In multi lingual environment where in a single image document have more than one script occur there is need of script identification system. Automatic identification of scripts in document facilitates (i)Automatic archiving of multilingual documents, (ii) Searching online archives of document images, (iii) Selection of script specific OCR in a multilingual environment. The main objective of this system is to identify the specific script and feed them into their specified Optical Character Recognition (OCR) system. OCR is the system which converts the image document into editable text document. Script identification of written text in the domain of Indian script based languages is a well-studied research field. In this paper a technique of script Identification is described to discriminate three major south Indian scripts: Oriya, Telugu and Kannada. These three scripts are member of Brahmi script and most of the character shapes are near similar. This method is applied over segmented line from the image document and it is completely free from size and font. The proposed technique uses the basic distinguishable features based on texture analysis. The approach is based on the analysis of horizontal projection and vertical projection profile. We obtain overall 98.64% accuracy from test dataset of three ancient mix document images at line level.

Keywords
OCR, Script Identification, KNN, Oriya, Telugu, Kannada, Projection profile.

1. INTRODUCTION
Identification of the script in a document image is of primary importance for multi-lingual documents. Documents can be sub-divided on the basis of the scripts and further, the script of the text in the page image applied to specific OCR system and extract textual information. The basic problem involved here is that of classification of the document regions on the basis of script using image based features. So, a pre-processor to the OCR system is necessary to identify the script type of the document, so that specific OCR tool can be selected. Among the earlier works in this area, M C Padmal et al proposed an approach of 8 discriminating features for Devanagari, Telugu and English script [1]. B.V. Dhandra et al proposed another approach for word level script identification in bilingual documents through discriminating features proposed using histogram [2]. Sukalpa Chanda et al proposed an approach of script identification using Gaussian Kernel SVM classification for Sinhala, Tamil and English script and get 94.01% accuracy for Tamil script [3]. U.Pal et al proposed an approach for script identification on several south Indian scripts with discriminating features and overall accuracy is 97.52% [4].
The general structure of OCR system takes input from script identification to convert image into editable text of appropriate script as shown in Fig 1.

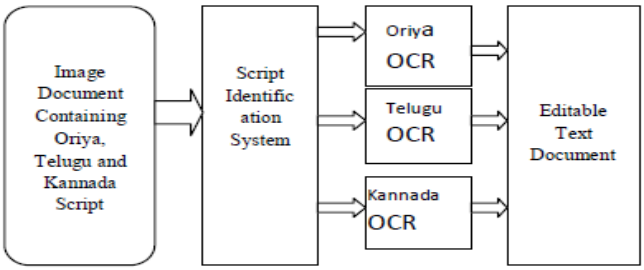


Fig1: Script identification as a part of OCR system

This paper deals with a line-wise script identification scheme for three popular south Indian scripts Oriya, Telugu and Kannada. Main objective of this system is to identify the specific script and feed them into their specified Optical Character Recognition (OCR) system. The K nearest neighbour is used in this approach for efficient classification on the base of extracted features. Approach methodology consists several phases from pre-processing to feature vector classification.

1.1 Properties of Script
Structure of Indian languages is more difficult than of European languages because of the large number of vowels, consonants, and conjuncts especially in south Indian scripts like Oriya, Telugu and Kannada. Oriya script consists of 13 vowels, 3 vowel modifiers, 37 consonants, 10 numerical digits and more than 59 composite characters. Out of 52 basic characters 37 characters have a convex shape at the upper part. This approach for script identification use the major feature of Oriya elementary characters is that most their upper one third is circular and vertical straight line at their rightmost part.

Telugu script is derived from Brahmi alphabet in 12th century. It consists of 60 letters with 16 vowels, 3 vowel modifiers and 41 consonants [5]. Telugu is one of the most complex scripts with highly curved letters and have no linear strokes. In most of the Telugu text line have tick like shape at their upper part. Kannada script has 48 characters, called varnamale. There are 14 vowels 34 consonants and 10 numerals. Consonants are divided into grouped consonants and ungrouped consonants. Vowels along with consonants constitute basic character. Vowel modifiers can appear to the right on the top or at the bottom of a base consonant. In most of the Kannada script text line horizontal stroke at their upper part.

2. PREPROCESSING
Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

2.1 Binarization

In pre-processing of the input image document system need two major steps binirization and line segmentation. One of the popular methods to binarize an image is thresholding approach because of its simple operation and efficient result.

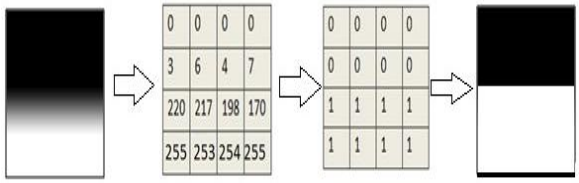


Fig 2: Thresholding Binarization Process

This approach gives good results in case of gray scale images. Gray scale image consists of intensity values between 0 to 255 which need to convert into 0 and 1 only according to their appropriate class [6]. System takes input image into gray tone (0-255) and using a thresh-holding approach converts them into two-tone images (0 and 1), black pixels having the value 0's correspond to object and white pixels having value 1's correspond to background as shown in Fig2.

2.2 Line Segmentation

In the *Text Line* segmentation in a document image processing represents a labelling process, which consists in assigning the same label to every individual unit. Projection method is most efficient for binary images as it simply divide the image into two categories either foreground or background. This method calculates sum of black and white pixel in every row or column of the image matrix and called projection profiles. Using this projection feature system partition an image into two regions foreground and background. Binary image consists of only two intensity value either 0 or 1 and here 0 uses for black pixel and 1 for white. 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 [7].

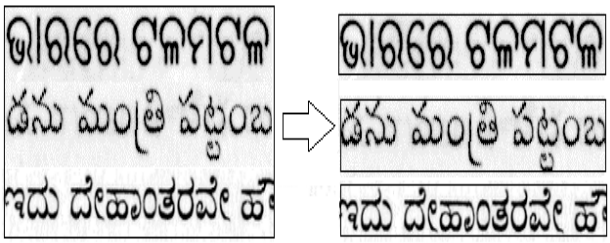


Fig 3: Text Line Segmentation

Using the same technique, the whole document is scanned and each line is detected and saved in a temporary array for further processing.

2.3 Resizing

In the proposed approach identification occur on line wise and input image text line may vary in size so to make system more efficient it needs to change the text size into a standard size. Bicubic interpolation technique is used to resize the text line as it gives more accurate result in binary images. In bicubic interpolation algorithm system simply change input image into a relative proportion to change into result image. This process considering the closest (4x4) neighbourhood of known pixels in which total 16 pixel are there. Since all the existing pixels in image are at different distance from unknown pixel and closer pixel are given higher priority in calculation.

Fig 4: Resizing of Text Line

3. FEATURE EXTRACTION

The proposed system has used horizontal projection profile and vertical projection profile based feature extraction technique to classify the scripts.

Feature1: Number of Vertical line equal to threshold height: It is the number vertical strokes in a single line which can be detect using vertical projection profile. If in a column number of black pixel is equal to the threshold height then it counts as a vertical stroke feature.

Fig 5: Vertical line in Oriya Script

Feature2: Upper profile component: Suppose each upper portion of the line is located within a rectangular boundary. The vertical distances from top side of the frame to the character edge are a group of parallel lines which can call here top profile. Most of the Oriya characters have transition points because of their concave shape. By transition mean change of the profiles from decreasing mode to increasing mode or vice-versa. In this feature detection most of the transition point changes into increasing mode and then into decreasing mode. Change in transition can be calculated on the basis of distance vector from bottom line to curve shape at top part of the text line. Theses distance vectors are in the pattern of increasing then decreasing mode detects the Upper profile component feature in Oriya script as shown in Fig 6.

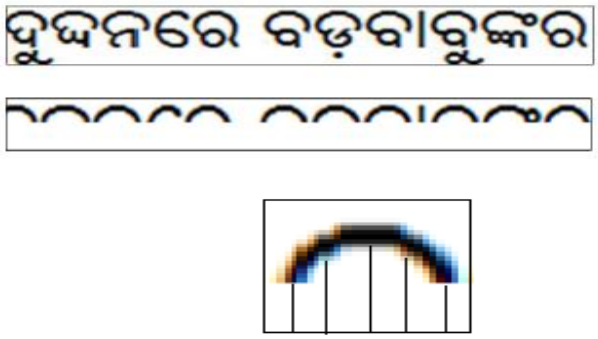


Fig 6: Concave shape in top profile of Oriya script

Feature3: Frequency count in top segment: To detect this feature first need to extract top profile of the line from where the horizontal projection value is maximum and after have to count the frequency of each distinct value of vertical projection. Due to maximum number of horizontal stroke present in top profile of Kannada script this feature identifies the language.

Feature4: Horizontal Stroke like component in top segment: To detect this feature first extract top profile of the line and segment it up to connected component. In each component at maximum horizontal projection value detects maximum vertical value at the end or just before the end. To extract this feature basic projection profiles are used. Kannada script can be easily detects using horizontal stroke in top profile.

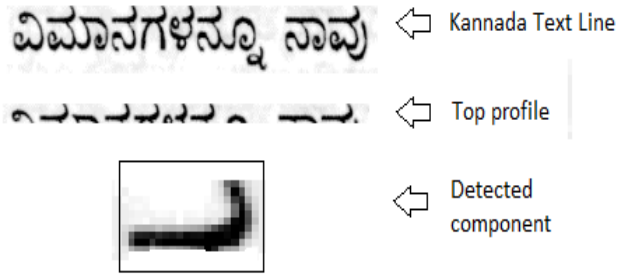


Fig7: Horizontal stroke like shape in top profile of Oriya script

Feature5 Tick mark like component: A component is said to have the shape of the tick like structure if the pixel values of the components are in the sequence as first x co-ordinates are increases with decrement in y co-ordinate again after a fixed point x co-ordinate are increases with increment in y co-ordinates.

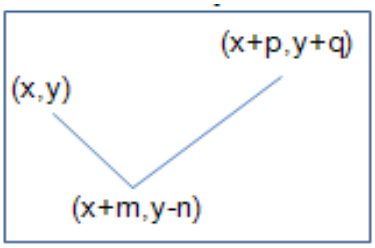


Fig7: Tick mark like shape in top profile of Telugu script

4. FEATURE CLASSIFICATION

The classification stage is the main decision making stage of a script identification system in which it uses the extracted feature as input to identify the text segment according to the pre-set rules.

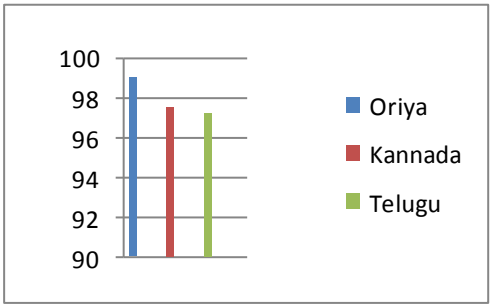
Table1: Features for Classification

	FEATURE	Oriya	Kannada	Telugu
F1	Vertical line	Yes	---	---
F2	Upper profile	Yes	---	---
F3	Frequency count	---	Yes	---
F4	Horizontal Stroke	---	Yes	---
F5	Tick mark	---	---	Yes

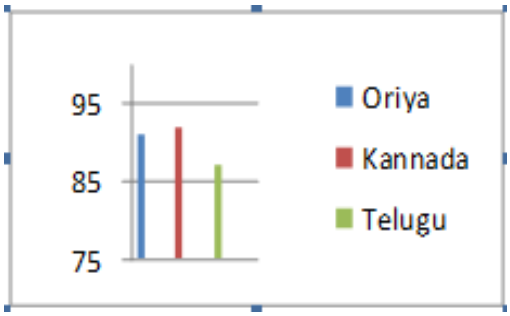
For separation of scripts in the system a range K-nearest neighbour classification approach is used in which all extracted features of the document analysed and propose a class for individual feature. In the first phase of KNN classification storing the feature vectors of the training data set consists of large number of image document of tri-lingual script.

5. EXPERIMENTAL RESULT

In this section we evaluate the proposed method of script identification over 500 image document from different sources. It contains approx 10000 text lines of three scripts which identifies. Documents are of different font size for getting more accuracy in classification. The images were scanned from newspaper, magazine, book, money order form, computer printouts, translation books etc. Each line contains at least 12 characters. Two graphs depict the testing result over 500 image document in which some are noisy and some noiseless.



a. Result for noiseless data



b. Result for noisy data

Fig 8: Result Graphs (a) Noiseless Data (b) Noisy Data

6. CONCLUSION

We compared our results with some of the recently published work on script identification over south Indian scripts. P.A. Vijaya et al [8] proposed a rule based classifier using top and bottom profile features is used and obtained 96.6% recognition accuracy. Comparative to other papers feature detection methods of our approach are more robust and giving more success rate.

In this paper, a simple and efficient algorithm for script identification of Kannada, Telugu, and Oriya text lines from printed documents is proposed. The approach is based on the analysis of horizontal projection profile and vertical projection. The system does not require any training data. The system exhibits an overall accuracy of 98.64%. The work could be extended to word level script identification and for all Indian scripts.

7. REFERENCES

[1] M C Padma and P A Vijay "Identification of Telugu Devnagri and English Script using discriminating feature "International Journal of Computer science & Information Technology (IJCSIT), Vol 1, pp. 64-78 , November 2009.

[2] Rajesh Gopakumar, N V Subbareddy, Krishnamoorthi Makkithaya, U Dinesh Acharya "Zone-based Structural feature extraction for Script Identification from Indian Documents" 2010 5th International Conference on Industrial and Information Systems, ICIIS 2010, 978-1-4244-6653-5/10/$26.00 ©2010 IEEE pp. 420-425 ,2010.

[3] B.V. Dhandra, Mallikarjun Hangarge, Ravindra Hegadil and V.S. Malemathl "Word Level Script Identification in Bilingual Documents through Discriminating Features" IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. Feb. 22-24, 2007. Pp.630-635.

[4] U. Pal, S. Sinha and B. B. Chaudhuri "Multi-Script Line identification from Indian Documents" Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)0-7695-1960-1/03 $17.00 © 2003 IEEE.

[5] P Nagabhushan, S.A. Angadi and B.S. Anami," An Intelligent Pin code Script Identification methodology based on texture analysis using modified invariant moments "In Preceding of ICCR-2005,pp.615-623.

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

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

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

[9] Sukalpa Chanda, Srikanta Pal and Umapada Pal," Word-wise Sinhala Tamil and English Script Identification using Gaussian Kernel SVM " In Preceding of IEEE-2008 978-1-4244-2175.