ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Script Identification: An Image Processing Approach

Ram D.Sarje[1], V. R. Ratnaparkhe [2]

P.G. Student, Department of EE, Government College of Engineering, Aurangabad, Maharashtra, India[1]

Assistant Professor, Department of EE, Government College of Engineering, Aurangabad, Maharashtra, India[2]

ABSTRACT: - Script identification is necessary in multi script environment. Most of documents have text in different languages and scripts. Script identification is also required to feed document with Optical character recognition tool. Museums have many stored images of old document. Document image analysis helps to process that old images and utilize to researchers. In this paper, identification of Devanagari, English and Telugu script from multi script environment. Horizontal line recognition and tick shape components are used for discrimination. Experiments conducted for 1800 text lines of different script. The overall efficiency is 98%.

KEYWORDS: Script Identification, Document Image Processing, Horizontal Line Detection and Tick shape component.

I. INTRODUCTION

In recent years, the physical documents can be converted into electronic documents to provide communication and storage of data. However, the use of physical documents is necessary. For instance, fax machine is used for communication purpose worldwide. Paper is very comfortable and transmission medium is secure. So, there is demand of physical documents from last years and it will continue in future also. Document image processing technique is used to extract, read and store information from physical document.

The task for document image analysis is automatic reading of text lines from document image. Optical recognition (OCR) tool performs this task. To select a particular OCR tool, the type of script should be priority known. Hence preprocessing is essential to identify the script.

In India, 18 regional languages are derived from 12 different scripts. A document page like bus reservation forms, question papers, language translation books and money-order forms may contain text lines in more than one script/language forms [2].One script could be used to write more than one languages. For example, in Devanagari script there are five different languages like Sanskrit, Marathi, Hindi, Nepali and Rajastani. With this context, in this paper the problem of language type of text is recognized.

II. LITERATURE SURVEY

Automatic script identification has been a challenging research problem in a multilingual environment over the last few years. All existing works on automatic language identification are classified into either local approach or global approach. Local approaches extract the features from a list of connected components like line, word and character in the document images. In contrast, global approaches employ analysis of regions comprising at least two lines and hence do not require fine segmentation. As a result, global approaches are faster since no segmentation of the document image into lines, words and characters is required [5].

A. Local approaches on Indian scripts:

Pal and Choudhuri group [2] is working on Indian script identification by extracting features from projection profile and water reservoir concept. Basavaraj Patil [4] has prepared a neural network based system for script identification of Kannada, Hindi and English languages.

B. Global Approaches on Indian Scripts:

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Monothetic separation of multi-lingual Document using Text line of different languages is done by M.C.Padma and P.A.Vijaya [1].Santanu Choudhuri, et al [3] have proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier.

III. DATA COLLECTION

Standard database of Indian documents are not available. Data base construction with respect to language identification problem may be complex. For that purpose font size and font type of each language needs to be fixed. In this paper, it is assumed that the input data set contains documents having the text lines of Telugu, Devanagari and English scripts. It is assumed that the language type, font and size of the text words within a text line are same. Two different data sets are used out of which one is manually constructed and second is considered from scanned document images. The document of Telugu script is created using the Microsoft word software and these text lines are imported to the Micro Soft Power point program. In the Micro Soft power point, a portion of the text lines was saved as black and white JPEJ image. The font type of Times New Roman is used for English language. The font sizes of 10 to 12 are used for English text lines. The input images of all script are constructed by clipping only text portion of the document downloaded from the Internet. The sample database of Telugu and Marathi languages are shown in Fig. 1

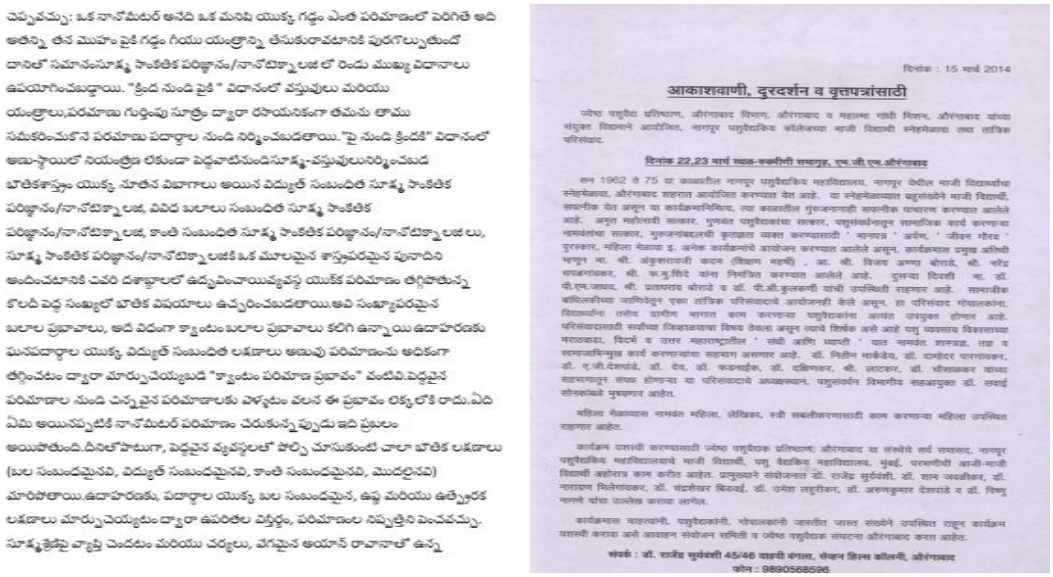

Fig. 1: Sample Images of Telugu and Marathi language.

A. Pre-processing:

Any language identification method requires good quality image input of the document, which implies that the document should be noise free and skew free. In this paper, the pre-processing techniques such as noise removal and skew correction are not necessary for the datasets that are manually constructed by downloading the documents from the Internet. However, for the datasets that are constructed from the scanned document images, pre-processing steps such as removal of non-text regions, skew-correction, noise removal and binarization is necessary. In this paper, text portion of the document image was separated from the non-text region manually.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

IV. PROPOSED MODEL

Each script considered a finite set of text patterns called alphabets. Grouping of alphabets of each script gives meaningful text information in the form of a word, a text line or a paragraph. Thus, script has different visual appearance. The different visual appearance of every script is due to the presence of the lines shape like horizontal lines, vertical lines, upward curves, downward curves and so on [1].
The presence of such segments in a particular script is used as visual clues to identify the type of even the unfamiliar script. It was motivated to adopt the idea of human visual perception capability into the proposed model to use the distinct features exhibited by each script [2]. So, the target of this paper is to identify the script type of the texts without reading the contents of the document.

A. Properties of Telugu and Devanagari Script:

Many characters of Devanagari script have a horizontal line at the upper part called headline which is named as siro-rekha in Devanagari. When two or more basic or compound characters are combined to form a word, the character headline segments mostly join one another and generate one long headline for each text word. These long horizontal lines are present at the top portion of the characters and they are used as supporting features in identifying Devanagari script. Using this feature Marathi text line could be strongly separated from English and Telugu languages.
Telugu is the official language of Andhra Pradesh. Telugu script is derived from Telugu script itself. Most of the Telugu characters have tick shaped structures at the top portion of their characters. Also, majority of Telugu characters have upward curves present at their bottom portion. These distinct properties of Telugu characters are helpful in separating them from Marathi and English languages.

B. Horizontal Line Detection:

Horizontal line detection algorithm is used to identify Devanagari script. First, convert input image into binary image. Then scan rows and columns up to (r) and (c-n) from initial pixel. Here ‘n’ is representing number of connected pixels. Also vary ‘n’ from 2:8. Compare pixel (i, j) to pixel (i, j+n).If this is equal then treat it as zero otherwise pixel (i, j) =1. Finally compute number of black pixel. If pixel (i, j) and pixel (i, j+n) are not equal means these pixels are not connected. So, there is no headline found in given language. For English and Telugu script there is no horizontal line so numbers of black pixels are zero.

Algorithm:
Input: Pre-processed text images of Devanagari, Telugu and English scripts.
Output: Range of black pixels values.
1. Scanning of image.
2. For each row find minimum 2 and maximum 8 horizontally connected pixels.
3. Then assign its value as 0.
4. Find number of black pixels

C. Tick Shape Component:

The upper part of Telugu script has tick shaped (√) structure. For detecting that tick shape pixel sequence are in (i, j), (i+1, j+1), (i+2,j+2), … , (i+k, j+l), (i+k-1, j+l+1), (i+k-2, j+l+2), (i+k-3, j+l+3), ... ,(i+k-k1,j+l+l1), where k1=i+k and l1>l. In this paper, the sum of tick shape component is equal to 7. Make that component equal to 1. So for pixel value greater than 7 or less than 7 there is no tick component. That pixel value component make zero. Calculate value of white pixel. For Marathi and English language such tick components are rarely found.

Algorithm:
Input: Pre-processed text images of Devanagari, Telugu and English scripts.
Output: Range of white pixels values.
1. Find pixels which gives tick shape as shown in Fig 2.
2. Then assign its value as 1.
3. Find number of white pixels.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

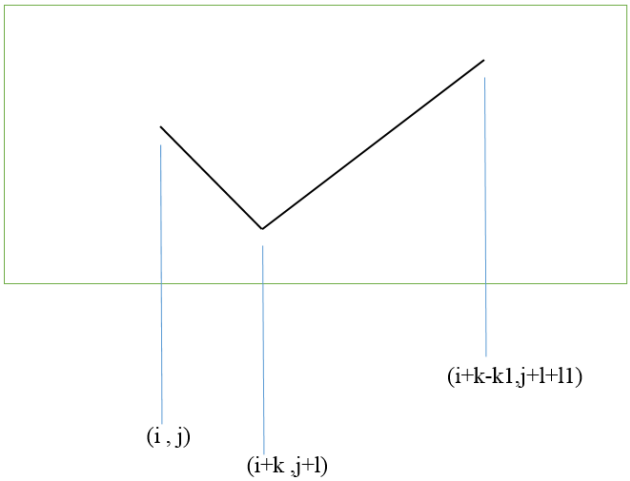

Fig. 2: Tick shape component in matrix

D. Flowchart
The overall flow of script identification method is shown in Fig. 3.

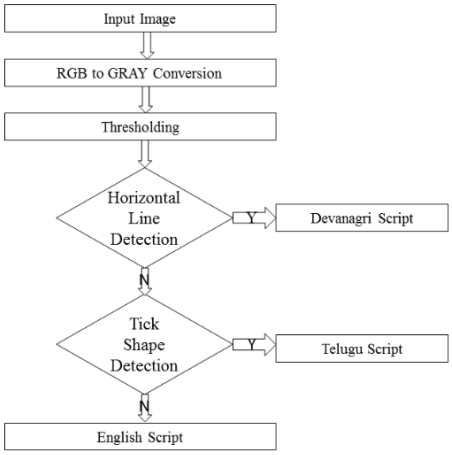

Fig. 3: Flowchart of Devanagari, Telugu and English script

V. RESULTS

For Telugu script, the tick shape components are separated from original image of Telugu language. The number of white pixels is calculated. If the number of tick components will be calculated from white pixels then divide total number of white pixel by 7 because most of tick components have that much of pixel length. The tick shape structure obtained from Telugu language from Telugu script are shown in Fig. 4.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

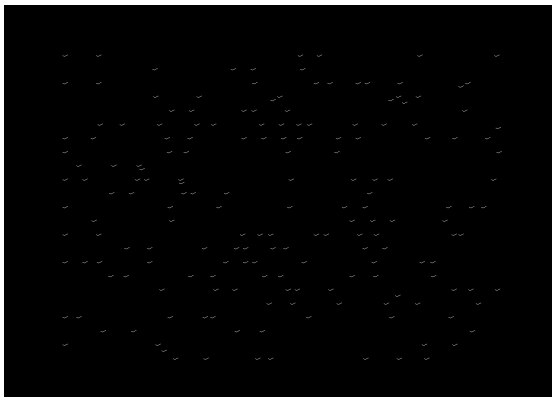

Fig. 4: Tick shape components detected for Telugu script.

The tabular result of tick component of Devanagari, Telugu and English script are calculated in Table 1. For Telugu script, tick components are more than other script.For other script, threshold is set according to value of pixels.

Table 1: Tick Detection Algorithm

Input	Class	No. of white pixels
Image 1	Devanagari	77
Image 2		91
Image 3		161
Image 1	Telugu	1281
Image 2		1225
Image 3		1253
Image 1	English	0
Image 2		0
Image 3		7

Also, for Devanagari script the horizontal line are detected. The total numbers of black pixels are calculated and these lines are shown in Fig. 5. Horizontal line pixels are calculated which is illustrated in Table 2. It is observed that Telugu and English script have no horizontal lines. So, Devanagari script is identified from other script.

Fig. 5: Horizontal line component obtained from Devanagari script.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Input	Class	No. of Black pixels
Image 1	Devanagari	2896
Image 2		2786
Image 3		2662
Image 1	Telugu	0
Image 2		0
Image 3		0
Image 1	English	0
Image 2		0
Image 3		0

VI. CONCLUSION

In this paper, different scripts are identified using text line of Devanagari, English and Telugu script. The method uses top profile of individual languages of different scripts, word segmentation is not necessary. A document may contain different languages within text line. The given method cannot differentiate scripts, this may be future scope to detect script at word level.

REFERENCES

[1] M. C. Padma, P. A. Vijaya, "Monothetic Separation of Telugu, Hindi and English Text Lines from a multi Script Document", IEEE international Conference on Systems, Man and Cybernetics, USA, October 2009.
[2] U. Pal, B. B. Choudhuri, "Script Line Separation from Indian Multi-Script Documents", 5th Int. Conference on Document Analysis and Recognition, 406-409, 1999.
[3] Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet, "Identification of Scripts of Indian Languages by Combining Trainable Classifiers", ICVGIP, Bangalore, India, Dec.20-22, 2000.
[4] S. Basavaraj Patil and N V Subbareddy, "Neural network based system for script identification in Indian documents", Sadhana Vol. 27, Part 1, pp. 83–97. 2002.
[5] Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy, "Script Identification from Indian Documents", LNCS 3872, pp.255-267, DAS 2.