A Novel Way of Identifying Telugu, Tamil and English Scripts by Priority check using Discerning Features

Srihareendra Bodduluri, Mani Krishna, Ratan Babu and M.V.Raghunadh
Department of Electronics and Communication National Institute Of Technology, Warangal, India

Abstract: *India is a vast country with more than 22 recognized languages and 12 major scripts .There are many Optical Character Recognition(OCR) systems available in the market but mostly for roman Chinese ,Japanese and Arabic characters .Not enough research is done in recognizing Indian scripts especially south Indian languages. India is a multilingual multi script country therefore it is necessary to identify different scripts regions of a document in order to recognize the characters of the individual documents. This papers main focus is to develop a model to identify Telugu, Tamil and English scripts from a printed documents. The proposed model uses distinctive features extracted in a priority check scheme. The accuracy of the scheme proposed is 99.33%.*

Key Words: *OCR, Script Identification, Distinctive features, priority check scheme*

I. Introduction

In the recent times there has been a rise in the digitalization of documents. However the physical documents are still in use. There is great demand for software which scans, stores separates and analyses the physical documents. Many published paper are available towards postal automation of non-Indian Language documents and sorting systems are also available for postal automation in several countries like USA, UK, Japan, Germany etc. But no such sorting system is available for Indian script. System development towards postal automation for a country like India is difficult because of its multi-lingual and multi-script behaviour.

In a multi-script multi-lingual country like India (India has 18 regional languages derived from12 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. One script could be used for writing more than one languages. For example, languages such as Hindi, Marathi, Rajasthani, Sanskrit and Nepali are written using the Devanagari script; Assamese and Bangla languages are written using the Bangla script. In order to reach a larger cross section of people, it is necessary that a document should be composed of text content in different languages. However, for a document having text information in different languages, it is necessary to pre- determine the language type of the document, before employing a particular OCR on them. With this context, in this paper, the problem of recognizing the language type of the text content is addressed. However, it is perhaps impossible to design a single recognizer, which can identify a large number of scripts/languages. As a via media, this paper proposes to work on the prioritized requirements of Andhra Pradesh and Tamil Nadu states in India. Both the languages have attained classical language status and are most spoken after Hindi and Bengali. .So, when it comes to automation, assuming that there are three OCRs for Telugu, Tamil and English languages, a pre-processor is necessary by which the language type of the different texts lines are identified. In this paper, a script identification technique to identify the text lines of Telugu, Tamil and English languages from a tri-lingual document is presented.

The rest of the paper is organized as follows, Section II deals with literary survey , Section III deals with pre-processing and data collection ,Section IV deals with proposed model and Section V deals with results and conclusions.

II. Literary Survey

Structural features can represent various global and local properties of characters with high tolerance to distortions and style variations. These features describe a pattern in terms of its topology and geometry by giving its global and local properties. Many different types of features have been identified by Suen et.al. [1] In the literature that may be used for character and numeral. Two main categories of those features are Structural (topological) and Global (statistical). Trier et al. [2] summarized a good survey on feature extraction method for character recognition. The author mentioned that different types of features can be extracted depending on the representation forms of characters, which can be grouped as grayscale images, Binary images, character contour and character skeletons. Heutte et al. [3] says that some of the main structural features include features like number and intersections between the character and straight lines, number of vertical and horizontal lines, holes position, end points, presence of loops, number of loops, number of intersections and junctions. These features are generally hand crafted by various authors for the kind of pattern to be classified.

Majority of the work on Indian script identification has been carried out by Pal, Choudhuri and their team [4, 5, 6and9]. Pal and Choudhuri [4] have proposed an automatic technique of separating the text lines from 12 Indian scripts (English, Devanagari, Bangla, Gujarati, Tamil, Kashmiri, Malayalam, Oriya, Punjabi, Telugu and Urdu) using ten triplets formed by grouping English and Devanagari with any one of the other scripts. This method works only when the triplet type of the document is known. Script identification technique explored by Pal [5] uses a binary tree classifier for 12 Indian scripts using a large set of features. The binary tree classifier seems to be complex since the features are extracted at line, word and even at character level. From the literature it is observed that adequate work has been carried out on bi-lingual and tri-lingual documents of Indian languages specifically with respect to some Indian states [7, 8, 9, 10, 11, 12 and 13]. Basavaraj Patil et. al. [7] have proposed a neural network based system for script identification of Kannada, Hindi and English languages. Word level script identification in bilingual documents through discriminating features has been developed by Dhandra et. al. [8].A method to automatically separate text lines of Roman, Devanagari and Telugu scripts has been proposed by Pal et. al [9]. Lijun Zhou et. al. [10] have developed a method for Bangla and English script identification based on the analysis of connected component profiles. Padma et.al. [11, 12] have proposed a method based on visual discriminating features to identify Kannada, Hindi and English text lines. Vipin Gupta et.al. [13] have presented a novel approach to automatically identify Kannada, Hindi and English languages using a set of features viz., cavity analysis, end point analysis, corner point analysis, line based analysis and Kannada base character analysis. Survey on the existing techniques of script identification of Indian documents shows that majority of the work is constrained to languages followed by a particular state.

III. Data Collection And Pre-processing

A. Data Collection

Currently in India there is no standard database for the documents. Data base construction with respect to the language identification problem seems to be complex since the factors like the font type and font size of each language needs to be considered. The document of English language was created using the Microsoft word software and these text lines were imported to the Micro Soft Paint program. In the Micro Soft Paint, a portion of the text lines was saved as black and white Bit Map (BMP) image. The font type of Times New Roman, Arial, Bookman Old Style and Tahoma were used for English language .Telugu script writing system is alpha syllabify in which all consonants have an inherent vowel. Other vowels are indicated with diacritics, which can appear above, below, before or after the consonants .Telugu has 16 vowels and 36 consonants. There are about 250 basic, modified and compound characters shapes in Telugu. Writing style is from left to right in a horizontal manner. Upper and lower case distinction is not present in Telugu. It is not like the English words which are only constructed by 26 letters. Figure 1 gives the complete character set of Telugu script. Data for Telugu language is collected from various sources like internet, scanned images of books, newspaper etc.

అ ఆ ఇ ఈ ఉ ఊ ఋ ౠ
ఎ ఏ ఐ ఒ ఓ ఔ అం అః
క ఖ గ ఘ ఙ
చ ఛ జ ఝ ఞ
ట ఠ డ ఢ ణ
త థ ద ధ న
ప ఫ బ భ మ
య ర ల వ ళ
శ ష స హ ఱ

Figure 1. Telugu characters set

There are 12 vowels, 18 consonants, 216 composite letters, one special character (AK) and 14 other characters in Tamil script Composite letters are not basic and they are derived by Combining consonants and vowels as described in [9]. We have identified some 67 Tamil characters as the basic characters (Vowels, Consonants, and composite letters) and if one recognizes these 67 characters then all the 247 characters can be recognized. Figure 2 gives complete Tamil consonant and vowel character set. To test the proposed model, two different data sets were constructed out of which one data set was constructed manually similar to the dataset constructed for training and the other data set was constructed from the scanned document images. The printed documents like text books and magazines were scanned through an optical scanner to obtain the document image. The HP Scanjet 5200c series scanner was used to obtain the digitized images. The scanning was performed in normal 100% view size at 300 dpi resolution. Manually constructed dataset is considered a good quality dataset and the data set constructed from the scanned document images are considered as poor quality data set. The test datasets were constructed such that 300 text lines from each of the three languages - Telugu, Tamil and English, were present from each of the good quality and poor quality datasets

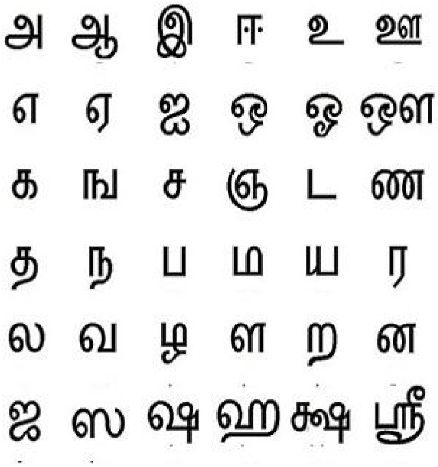


Figure 2. Tamil characters set

B. *Pre-processing*

The sequences of pre-processing steps are as follows

(a) Noise Removal

Noise is defined as any degradation in the image due to external disturbance. Quality of documents depends on various factors including quality of paper, aging of documents, quality of pen, colour of ink etc. Some examples of noise are salt and pepper noise, Gaussian noise. These noises can be removed to certain extent using filtering technique.

(b) ***Binarization***

Binarization is used to convert gray to binary images. It separates the foreground information. This is an essential part as we have objects of interest from other part of the common method employed in binarization threshold for the intensity of the image and the intensity values above the threshold to one all intensity values .Below the threshold are chosen intensity zero. Binarization is performed globally and locally. Figure
3a, 3b illustrates the original and binarized images

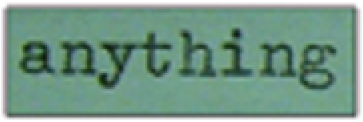


Figure 3a. Original image

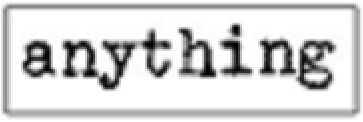


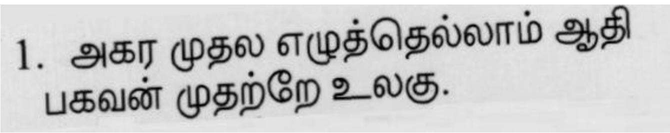
Figure 3b. Binarized image

(c)Skeletonization

Skeltonization is an image pre-processing operation performed to make the image crisper by reducing the binary valued image regions to lines that approximate the skeletons of the region.

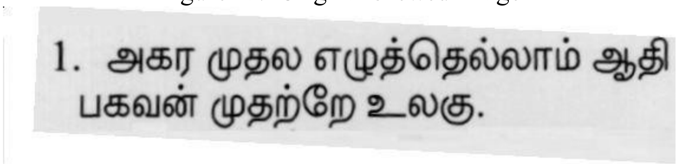
(d) Skew Detection and Correction

Casual use of the scanner may lead to a skew in the document image. Skew angle is the angle that the text lines of a recorded digital document make with the horizontal direction. This may cause problems in segmenting the image to extract its layout structure. Skew detection in Telugu text documents is made complicated by the presence of the vowel and consonant modifiers above and below the modified characters. Because of these, symbols belonging to a line of text do not fall on the same horizontal line. Further, there is no top bar as in other Indian scripts, like Devanagari, to aid the skew detection process. Figure 4a, 4b, 4c, 4d illustrates the skew in the scanned images



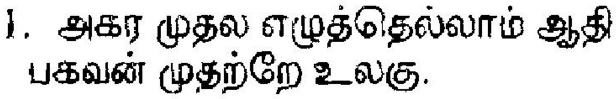
1. அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.

Figure 4a. Original skewed image



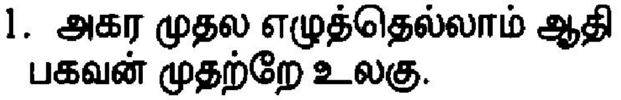
1. அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.

Figure 4b. Skew correction performed on binary image



1. அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.

Figure 4c. Skew corrected grey scale image



1. அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.

Figure 4d. Skew corrected binarised image

There exist a wide variety of skew detection algorithms based on the projection profile, Hough transform, line correlation, etc. The algorithm we are going to use is a combination of the profile method and the Hough transform method.

IV. Proposed Model

The scripts of the above mentioned languages have their own set of alphabets. Alphabets of one script are grouped together giving meaningful text information in the form of a word, a text line or paragraph. Thus, when the alphabets of the same script are combined together to yield meaningful text information, the text portion of the individual script exhibits a distinct visual appearance. The distinct visual appearance of every script is due to the presence of the segments like – horizontal lines, vertical lines, upward curves, downward curves, descendants and so on. The presence of such segments in a particular script is used as visual clues for a human 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. So, the target of this paper is to identify the script type of the texts without reading the contents of the document. The identification process is divided into two stages

The priority check algorithm is used for separation of lines or paragraphs in a given multilingual documents .Each line is divided into 3 zones top, middle and bottom zone as shown in the figure 5.

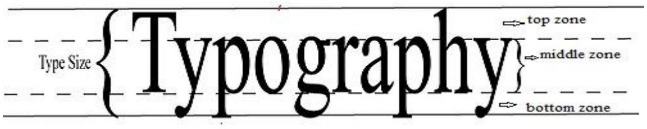


Figure 5. Zonal segmentation of the text

A. Features identified for Telugu script

The following signs are identified in the given document for Telugu language are shown in figure 6.

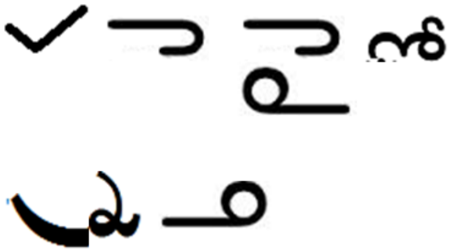
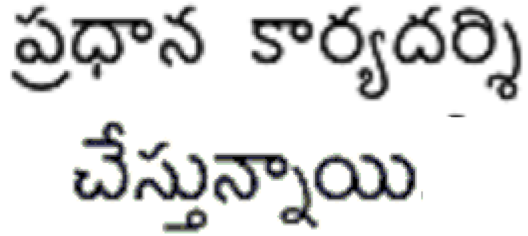


Figure 6. Assignment of signs with priority

In a given document the top zone and bottom zone are observed for the above signs with priority order as shown in the above figure 6.In a given line if a high priority symbol is observed the search shifts to next line else it shifts for searching the next priority.
Figure 7 and figure 8 illustrates the zoning.



ప్రధాన కార్యదర్శి

చేస్తున్నాయి

Figure7. Sample document in Telugu

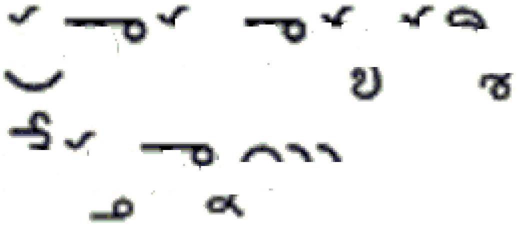


Figure 8. Identifying top and bottom zones of the sample document.

B. Features identified for Tamil script

The following signs are identified in the given document for Tamil language. Upper arc, upper arc with loop and presence of dot .The upper arc and dot are observed in the top zone .loop is observed in between the middle and bottom zone

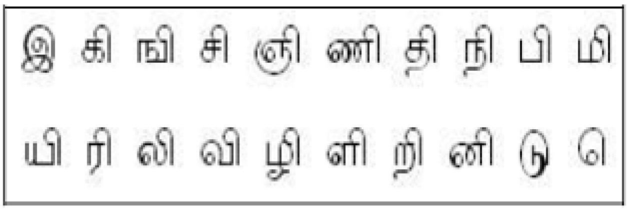


Figure 9a. Characters with upper arc

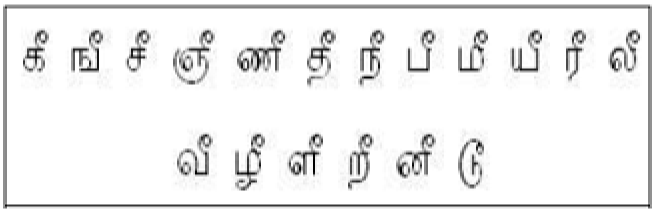


Figure 9b. Characters with upper arc and loop

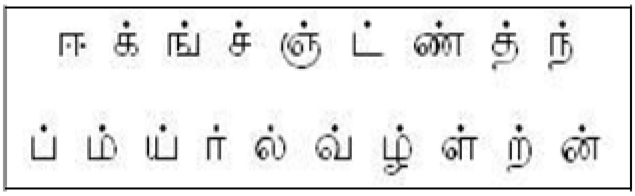


Figure 9c. Characters with dot

C. *Features identified for English script*

Telugu and Tamil scripts are devoid of slanting and cross lines. Making use of this property we are going to find slope ('θ') of a given text .so if 'θ' lies between -70° to +70°˒ we can assume the presence of a slanted line.

We can also make use of horizontal profile for identification of English language. The horizontal profile of English script gives two peaks one in the above half and the latter in below half. The peaks are of similar length, this is because of regularity in the English script. Figure 10 illustrates horizontal projection for a English sentence.

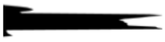
ENGLISH SENTENCE	HORIZANTAL PROJECTION
TIME GOES SO FAST PLEASE GO IN AND OUT OF YOURLIFE	

Figure 10. Illustrating horizontal projection in English

V. Results And Conclusions

The proposed system has been tested on over 1100 various types of documents and 2000 document images for each script The percentage of recognition is given in the below table.

Table 1: Percentage of Recognition data set (manually created) From the experimentations on the test data set, the overall

Input/output	Telugu	Tamil	English	rejected
Telugu	99.5%	------	0.3%	0.2%
Tamil	0.3%	99.25%	------	0.45%
English	0%	0.25%	99.25%	0.5%

accuracy of the system has turned out to be 99.33%. From the Table 1, it could be observed that the 99.5% accuracy is obtained for Telugu script. This is because of distinct features of the Telugu script. From the experimental observations, it is noticed that the recognition rate is 99.5% for Telugu script even for the text lines having only one word, whereas the recognition rate falls down for the text lines with one or two text words. The proposed algorithm is implemented using Mat lab R2010b. The average time taken to identify the script type of the document is 0.02986 seconds on a Pentium-IV with 2048 MB RAM based machine running at 2.4GHz. The approach of the paper is based on analysis of zones which is very fast in nature. This algorithm can be used for quick sorting of documents. Our future work is to improve the algorithm for word identification without segmentation

Acknowledgement

We would like to acknowledge our mentor, Sri M.V.Raghunadh, withoutwhose help and constant guidance this paper would not have been successfully completed.

References

[1] C. Y. Suen, M. Berthod and S. Mori, "Automatic Recognition of hand printed characters- the state of the Art", Proceedings of the IEEE, Vol. 68(4), pp. 469-487, 1980.*)*

[2] O.D.Trier, A.K.Jain and T.Taxt, "Feature Extraction Methods of Character Recognition: - a survey", Pattern Recognition, Vol.29 (4), PP. 641-662, 1996.

[3] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier and C.Olivier, "A structural/statistical feature based vector for handwritten character recognition", Pattern Recognition Letters, Vol. 19(7), pp. 629-641, 1998.

[4] U.Pal, B.B.Choudhuri, : Script Line Separation From Indian Multi-Script Documents, 5th Int.Conference on Document Analysis and Recognition (IEEE Comput. Soc. Press), 406-409,(1999).

[5] U. Pal, S. Sinha and B. B. Chaudhuri : Multi-Script Line identification from Indian Documents, In Proceedings of the Seventh International Conference on Document Analysis and Recognition(ICDAR 2003) 0-7695-1960-1/03 © 2003 IEEE, vol.2, pp.880-884, (2003).

[6] Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy, Script Identification from Indian Documents, LNCS 3872, pp. 255-267, DAS (2006).

[7] 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. © Printed in India, (2002).

[8] B.V. Dhandra, Mallikarjun Hangarge, Ravindra Hegadi and V.S. Malemath,: Word Level Script Identification in Bilingual Documents through Discriminating Features, IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. Pp.630-635. (2007).

[9] U. Pal and B. B. Chaudhuri, "Automatic separation of Roman, Devanagari and Telugu script lines", Advances in Pattern Recognition and Digital techniques, pp. 447-451, 1999.

[10] Lijun Zhou, Yue Lu and Chew Lim Tan,: Bangla/English Script Identification Based on Analysis of Connected Component Profiles, in proc. 7th DAS, pp. 243-254, (2006).

[11] M. C. Padma and P.Nagabhushan,: Identification and separation of text words of Karnataka,Hindi and English languages through discriminating features, in proc. of Second National Conference on Document Analysis and Recognition, Karnataka, India, pp. 252-260, (2003).

[12] M. C. Padma and P.A.Vijaya,: Language Identification of Kannada, Hindi and English Text Words Through Visual Discriminating Features, International Journal of Computational Intelligence Systems (IJCIS), Volume 1, Issue 2, pp. 116-126, (2008).

[13] Vipin Gupta, G.N. Rathna, K.R. Ramakrishnan,: A Novel Approach to Automatic Identification of Kannada, English and Hindi Words from a Trilingual Document, Int. conf. on Signal and Image Processing, Hubli, pp. 561-566, (2006).