RESEARCH ARTICLE OPEN ACCESS

Language Identification from a Tri-lingual Printed Document: A Simple Approach

Arjun M. Y.[1], Chirag H G[2]
[1]Dept. of C S & Engg., PES College of Engineering, Mandya-571401, Karnataka, India
[2]Dept. of C S & Engg., PES College of Engineering, Mandya-571401, Karnataka, India

Abstract—
In a multi-script, multilingual country like India, a document may contain text lines in more than one script/language forms. For such a multi-script environment, multilingual Optical Character Recognition (OCR) system is needed to read the multi-script documents. To make a multilingual OCR system successful, it is necessary to identify different script regions of the document before feeding the document to the OCRs of individual language. With this context, this paper proposes to work on the prioritized requirements of a particular region- Andra Pradesh, a state in India, where any document including official ones, would contain the text in three languages-Telugu, Hindi and English. So, the objective of this paper is to develop a system that should aim to accurately identify and separate Telugu, Hindi and English text lines from a printed multilingual document and also to group the portion of the document in other than these three languages into a separate category OTHERS. The proposed method is developed by thoroughly understanding the nature of top and bottom profiles of the printed text lines. Experimentation conducted involved 900 text lines for learning and 900 text lines for testing. The performance has turned out to be 95.67%.

Keywords- Document Image Processing, Multi-lingual document, Language Identification, Top Profile, Bottom Profile, Feature extraction.

I. INTRODUCTION

Automatic language identification plays an important role in processing large volumes of document images, particularly for a multilingual OCR system. In addition, the ability to reliably identify the language type using the least amount of textual data is essential when dealing with document pages that contain multiple languages. An automatic language identification scheme is useful to (i) sort document images, (ii) to select specific OCRs and (iii) search online archives of document image for those containing a particular language.

In a multi-script multi-lingual country like India (India has 18 regional languages derived from 12 different scripts [1]), a document page like bus reservation forms, question papers, bank challen, language translation books and money-order forms may contain text lines in more than one script/language forms. Under the three language formulae [1], adopted by most of the Indian states, the document in a state may be printed in its respective official regional language, the national language Hindi and also in English. Accordingly, a document produced in Andra Pradesh, a state in India, may be printed in its official regional language Telugu, national language Hindi and also in English. Further there is a growing demand for automatically processing the documents in every state in India including Andra Pradesh. With this context, this paper aims at identifying Telugu, Hindi and English languages specifically pertaining to the documents from Andra Pradesh, a state in India.

In the context of Indian language document analysis, major literature is due to Pal and Choudhuri [1]. This group worked on automatic separation of words from multi-script documents by extracting the features from projection profile and water reservoir concepts. Tan [2] has developed rotation invariant texture feature extraction method for automatic script identification for six languages: Chinese, Greek, English, Russian, Persian and Malayalam. Pal and Choudhuri [3] have proposed an automatic technique of separating the text lines from 12 Indian scripts (English, Devanagari, Bangla, Gujarati, Kannada, Kashmiri, Malayalam, Oriya, Punjabi, Malayalam, Telugu and Urdu) using ten triplets formed by grouping English and Devanagari with any one of the other scripts. Santanu Choudhuri, et al. [4] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. Chanda and Pal [5] have proposed an automatic technique for word wise identification of Devnagari, English and Urdu scripts from a single document. Gopal Datt Joshi, et. al. [6] has proposed script

Identification from Indian Documents. Word level script identification in bilingual documents through discriminating features has been developed by B V Dhandra et. Al. [7]. Neural network based system for script identification (Kannada, Hindi and English) of Indian documents is proposed by Basavaraj Patil et. Al. [8]. Lijun Zhou et. Al. [9] has developed a method for Bangla and English script identification based on the analysis of connected component profiles. In our earlier work, [10, 11] visual clues based methods for language identification of Telugu, Hindi and English text lines were developed. In this paper, we have proposed a method to identify Telugu, Hindi and English text lines using the features from the top and bottom profiles.

The concept of top and bottom profiles for a connected component is proposed by Lijun Zhou et. Al. [9]. However, in [9] only one feature was used to identify the two languages Bangla and English, where the visual appearance of the characters of these two languages is distinct. However, in the context of a multilingual country like India, the documents produced in every state are of tri-lingual type (a document having three languages). For such trilingual documents, the method proposed in [9] fails to identify the type of the language. With this backdrop, in this paper, we have proposed a technique that can identify the three languages Telugu, Hindi and English.

This paper is organized as follows. The section II describes the proposed technique of language identification. The details of the experiments conducted and the states of results obtained are presented in section III. Conclusions are given in section IV.

II. THE NEW MODEL

The new model is inspired by a simple observation that every script/language defines a finite set of text patterns, each having a distinct visual appearance which serves as useful visual clues to recognize the language [11]. The character shape descriptors take into account any feature that appears to be distinct for the language [11] and hence every language could be identified based on its discriminating features. The technical phrases top profile and bottom profile used in this paper are defined below:

Top-profile and Bottom-profile: The top-profile (bottom-profile) of a text line represents a set of black pixels obtained by scanning each column of the text line from top (bottom) until it reaches a first black pixel. Thus, a component of width N gets N such pixels.

The top-profile and bottom-profile of Telugu, Hindi and English text lines are shown in Figure 1.

A. Some Useful Discriminating Features in the characters of Telugu, Hindi and English text lines

It is observed that the most of the English characters are symmetric and regular in the pixel distribution. This uniform distribution of the pixels of English characters results in the density of the top profile to be almost same as the density of the bottom profile. However, such uniformity found in pixel distribution of the top and bottom profiles of an English text line is not found in the other two anticipated languages Telugu and Hindi. Thus, this characteristic attribute is used as a supporting feature to separate an English text line.

In Hindi, it is noted that a horizontal line called head-line found at the top portion of many characters are joined together in a word, generating a longer headline. Most of the pixels of these headlines become the pixels of top profile. This kind of line however is absent in the lower part. So for a Hindi text line, the density of top profile is comparatively much more than the density of the top profile of Telugu and English text lines. Thus density of the top profile of a Hindi text line projects the discriminating feature value to separate from Telugu and English text line. Another strong feature that could be noticed in a Hindi text line is that most of the pixels of the head line become the pixels of bottom profile, resulting in the density of both top and bottom profiles of a Hindi text line appearing at the position. However this distinct feature is absent in both Telugu and English text lines where the density top and bottom profiles occur at different positions. Using these two features Hindi text line could be strongly separated among the three anticipated languages.

It can be seen that, most of the Telugu characters have a tick-shaped structure and horizontal line-like structures towards the top portion of their characters. Also the bottom portions of the Telugu characters are curved in nature that reduces the compactness of the bottom portion of the text line. Hence for a Telugu text line, the density of the bottom profile is comparatively lesser than the density of the top profile. This makes the density of the top region higher than the density of the bottom region, which lies in between the range of the density of English and Hindi text lines. Thus the density of the top and bottom profiles of Telugu, Hindi and English text lines lie in distinct range. Hence the necessary features are extracted from the top and bottom profiles of the anticipated languages and are used to separate the three languages from a multilingual document.

B. Feature Extraction from Top and Bottom Profiles

The features used in the proposed technique are extracted from the top and bottom profiles as explained below:

Feature 1: Top-max-row: The attribute top-max-row represents the row of the top profile with maximum density i.e., the row with maximum number of black pixels (black pixels having the value 0's correspond to object and white pixels having value 1's correspond to background).

Feature 2: Bottom-max-row: The attribute bottom-max-row represents the row of the bottom profile with maximum density i.e., the row with maximum number of black pixels.

Feature 3: Top-max-row-no: The feature top-max-row-no (bottom-max-row-no) represents the row number at which the top-max-row (bottom-max-row) of top (bottom) profile lies. So the feature top-max-row-no (bottom-max-row-no) represents the row number of the top (bottom) profile at which the maximum number of black pixels lies.

Feature 4: Bottom-max-row-no: The feature bottom-max-row-no represents the row number at which the bottom-max-row of bottom profile lies. So the feature bottom-max-row-no represents the row number of the bottom profile at which the maximum number of black pixels lies.

Feature 5: Coeff-profile feature: The attribute coeff-top (coeff-bot) represents coefficient of variation of the top (bottom) profile and they are computed by using the equations (1) and (2) respectively.

$$\text{Coeff-top}=\sigma(\text{top-vector})/\mu(\text{top-vector})*100 \quad (1)$$

$$\text{Coeff-bot}=\sigma(\text{bottom-vector})/\mu(\text{bottom-vector})*100 \quad (2)$$

where σ and μ represents the standard deviation and mean of the corresponding vector. Then the feature coeff-profile is obtained using the equation (3).

$$\text{Coeff-profile}=\text{coeff-top}/\text{coeff-bot} \quad (3)$$

Then the percentage of the presence of these components is obtained through a training data set of 300 text lines from each of the three languages and it is given in Table 1.

TABLE I. RANGE OF FIVE FEATURE VALUES (F1:FEATURE 1- TOP-MAX-ROW; F2:FEATURE 2- TOP-MAX-ROW-NO; F3;FEATURE 3- BOT-MAX-ROW; F4;FEATURE 4- BOT-MAX-ROW-NO; F5:FEATURE 5 – COEFF-PROFILE).

	Telugu	**Hindi**	**English**
F1	38% to 55%	58% to 80%	34% to 40%
F2	7 to 9	10 to 11	11 to 12
F3	24% to 30%	25% to 32%	33% to 42%
F4	31 to 33	11 to 14	25 to 27
F5	1.5232 to 3.6573	0.2655 to 1.3852	1.6637 to 2.4051

C. Proposed Algorithm

The input document images are obtained by downloading the images from the Internet and hence do not require preprocessing such as noise removal and skew correction.

Step 1: Preprocessing: (i)The input document image is segmented into several text lines as explained in our previous paper [41]. (ii) A bounding box is fixed by finding the leftmost, rightmost, topmost and bottommost black pixel of each text line. (iii) The bounded text line is resized to 40X600 pixels.

Step 2 Learning algorithm: (i) Get the top-profile and the bottom-profile of each text lines. (ii) Get the top-max-row and bottom-max-row from the top and bottom profiles of each text lines. Get the range of the values of the necessary features (i) top-max-row value, (ii) top-max-row-no, (iii) bot-max-row, (iv) bottom-max-row-no, (v) coeff-profile value using a training data set of 300 text lines from each of the three languages and store those feature values in a knowledge base.

Step-3 Recognition algorithm: (i) The given new text line is preprocessed as explained before. (ii) The top and bottom profiles of the test text line are obtained. (iii) The values of the five features are computed from the top and bottom profiles. (iv) The five feature values of the test images are compared with the values of the knowledge base and a rule based classifier is used to classify the new text line to the type of the language that falls within that range.

OTHERS class: If the referenced text line is of the language type other than Telugu, Hindi and English, then such text lines could be grouped into a separate category called OTHERS, without identifying the type of the language as our main aim is to identify and select only Telugu, Hindi and English text lines.

III. EXPERIMENTAL RESULTS

The size of the sample image considered was 600x600 pixels. The system is trained to learn the behavior of the top and bottom profiles with a training data set of 300 text lines from each of Telugu, Hindi and English languages. The system is tested with a test data set of 900 text lines, having 300 text lines from each of the three languages. Sample output images of Telugu, Hindi and English text lines are shown in Figure-1, 2 and 3 respectively. Details of results obtained through extensive experimentation are given in Table-II. Figure II depicts the performance of recognition for a test data set of 800 text lines. From the experimental observation, we have noticed that high accuracy rate is achieved when the font type and font size of the test image is same as that of images used in training data set. We have found that 100% accuracy is obtained for English text lines with only uppercase letters. Good accuracy is obtained if the size of the text line is more than the size of the images used in training data set i.e., 40X600 pixels.

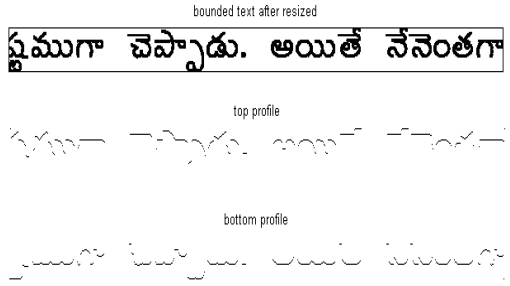

Figure 1. Sample output image of Telugu text line with its top and bottom profile.

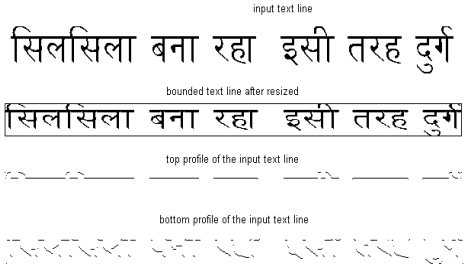

Figure 2. Sample output image of Hindi text line with its top and bottom profile.

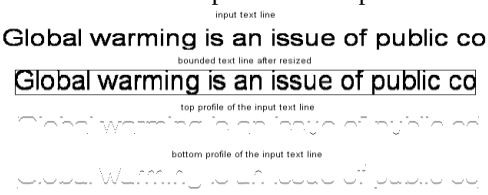

Figure 3. Sample output image of English text line with its top and bottom profile.

TABLE II. PERCENTAGE OF EXPERIMENTAL RESULTS.

	Telugu	Hindi	English	OTHERS
Telugu	94%	0%	1%	5%
Hindi	0%	98%	0%	2%
English	1%	0%	95%	4%

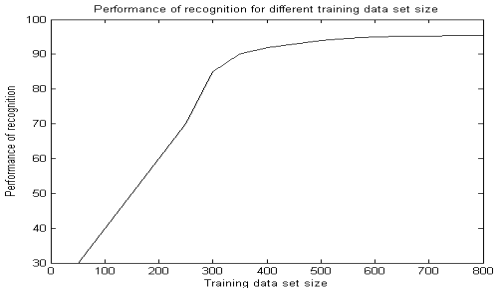

Figure 4. Percentage of Recognition of Telugu, Hindi and English text lines.

IV. CONCLUSION

In this paper, an algorithm for language identification of Telugu, Hindi and English text lines from printed documents is proposed. The approach is based on the analysis of the top and bottom profiles of individual text lines and hence does not require any character or word segmentation. Experimental results demonstrate that relatively simple technique can reach a high accuracy level for identifying the text lines of Telugu, Hindi and English languages. Our further research will focus on to improve the algorithm considering different font type and size and also to work on handwritten documents.

REFERENCES

[1] 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 © 2003 IEEE (vol.2, pp.880-884, 2003).

[2] T.N.Tan, "*Rotation Invariant Texture Features and their use in Automatic Script Identification*", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pp. 751-756, July 1998.

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

[4] Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet, "*Identification of Scripts of Indian Languages by Combining Trainable*

Classifiers", ICVGIP 2000, Dec.20-22, Bangalore, India.

[5] S.Chanda, U.Pal, English, Devanagari and Urdu Text Identification, Proc. International Conference on Document Analysis and Recognition, 2005, 538-545.

[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, February 2002, pp. 83–97. © Printed in India

[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. Feb. 22-24, 2007. pp.630-635.

[9] 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.

[10] 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, 2003, pp. 252-260.

[11] 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.

[12] Rafael C. Gonzalez, Richard E. Woods and Steven L. Eddins, "Digital Image Processing using MATLAB", Pearson Education, 2004.