# Language Identification from an Indian Multilingual Document Using Profile Features

M.C. Padma
Dept. of C S & Engg.,
PES College of Engineering,
Mandya-571401,Karnataka, India,
Email: padmapes@gmail.com

P. A. Vijaya
Dept. of E & C Engg.,
Malnad College of Engineering,
Hassan-573201, Karnataka,India
Email: pavmkv@yahoo.co.in

P. Nagabhushan
Dept. of Studies,
University of Mysore
Mysore,Karnataka, India
Email: pnagabhushan@hotmail.com

Abstract— **In order to reach a larger cross section of people, it is necessary that a document should be composed of text contents in different languages. But on the other hand, this causes practical difficulty in *OCRing* such a document, because the language type of the text should be pre-determined, before employing a particular OCR. In this research work, this problem of recognizing the language 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, in this research we have proposed to work on the prioritized requirements of a particular region, for instance in Karnataka state in India, generally any document including official ones, would contain the text in three languages-English-the language of general importance, Hindi-the language of National importance and Kannada –the language of State/Regional importance. We have proposed to learn identifying the language of the text by thoroughly understanding the nature of top and bottom profiles of the printed text lines in these three languages. Experimentation conducted involved 800 text lines for learning and 600 text lines for testing. The performance has turned out to be 95.4%.**

***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, in Karnataka, a state in India, generally any document including official ones would contain the text in three languages-English-the language of general importance, Hindi-the language of National importance and Kannada –the language of State/Regional importance. Further there is a growing demand for automatically processing the documents in every state in India including Karnataka. With this context, this paper focuses on identifying the language type of a document containing only these three languages Kannada, Hindi and English.

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, Tamil, 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

978-0-7695-3569-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICCAE.2009.35

script Identification from Indian Documents. Word level script identification in bilingual documents through discriminating features has been developed by 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. Our earlier methods [10, 11] focuses on language identification of Indian documents with a restriction that an input document should contain the text lines in regular font type only. So the method [10, 11] fails to identify the documents having italic font types. So, in this paper, we have proposed a method, which can handle documents having text lines in both regular and italic font types.

All existing language identification techniques including the methods mentioned above, fall into either local geometric or a global statistical approach. The local approaches analyze a list of connected components (like line, word and character) in the document images to identify the type of the language. However, these components are available only after line, word and character- (LWC) segmentation of the underlying document image. In contrast, global approaches employ analysis of regions comprising at least two lines and hence do not require fine segmentation. Consequently, the language classification task is simplified and performed faster with the global rather than the local approach. However, in practice it is not possible to apply the global approach for the types of documents where one paragraph or one line itself is composed of more than one language. For such types of documents where the language type differs at paragraph and/or line level, it is necessary to apply local approaches. So, in this paper, an attempt has been made to identify the three languages using local approach.

The concept of top and bottom profiles for a connected component is proposed by Lijun Zhou et. Al. [9]. However, in [9] the complexity lies in the feature extraction technique, since the feature (threshold) value is computed by getting the ratio of the sum of the differences of the black pixels of the top and bottom profiles. In addition, the two languages (Bangla and English) where the visual appearance and the structural form of the characters are very much distinct are considered. However, in the context of multilingual country like India, where every Indian states follow the three-language formula, the documents are printed their regional language, the national language Hindi and also in English. If such trilingual documents are considered, then the feature extraction method proposed in [9] fails to identify. With this backdrop, in this paper, we have proposed a technique that can handle the three languages Kannada, Hindi and English documents, restricting to concentrate on the documents from Karnataka state.

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. PROPOSED TECHNIQUE

Every language defines a finite set of text patterns and hence exhibits its own distinguishing features. The proposed approach is based on the characteristic features of the top and bottom profiles of the input text lines. The method does not place any emphasis on the information provided by individual characters themselves and hence does not require any character or word segmentation. The top-profile and bottom-profile of the input text line are computed as follows:

The top-profile (bottom-profile) of a text line is obtained by scanning each column of the text line from top (bottom) until it reaches a black pixel. Thus, for a component of width N, we get N such pixels. The top-profile and bottom-profile of Kannada, Hindi and English text line are shown in Figure 1, 2 and 3 respectively.

Choosing appropriate features useful for discriminating the different text lines of a multilingual document is an important step. The features used in the proposed technique are chosen with the following considerations: (i) Easy to detect; (ii) Feasible for identification; (iii) Accuracy; (iv) Speed of computation and (v) Independence of font and font size.

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. This characteristic attribute is used as a supporting feature to identify an English text line.

In Hindi, it is noted that many characters of these alphabets have a horizontal line at the upper part, which is called the headline. When two or more characters sit side by side to form a word, the headline portions touch one another and generate a long headline. Most of the pixels of these headlines are the pixels of top profile. This kind of line however is absent in the lower part, which leads to the distinction between the density of the top profile to the density of the bottom profiles. So for a Hindi text line, the density of top profile is much more than the density of the bottom profile. This characteristic feature is used to separate a Hindi text line.

It can be seen that, most of the Kannada characters have a horizontal line at the upper part and hence the pixels of these horizontal lines are the pixels of the top profile whereas such horizontal lines are absent in the bottom portion of the Kannada characters and hence the density at the bottom portion of the Kannada characters is comparatively less. Thus the nature of the top and bottom profile of the three languages Kannada, Hindi and English is found to be distinct and this method aims at extracting the features from the top and bottom profiles.

### *A. Feature Extraction from Top and Bottom Profiles*

To extract the required features from an input text line, it is necessary to consider the size of an input text line such that the input text line should possess the distinguishing features exhibited by the specific language. Through experimentation, it is decided to obtain the input text line

which should possess at least four text words and text line is normalized to a standard size of 40X600 pixels. The leftmost pixel, rightmost pixel, topmost pixel and bottommost pixel are obtained from the normalized text line and a bounding box is fixed. Then the top and bottom profiles are obtained for the bounded text line. From the top and bottom-profile, the row having maximum density i.e., the row with maximum number of black pixels (black pixels having value 0's correspond to object and white pixels having value 1's correspond to background) is selected. The value of the maximum number of black pixels present in a row of both top and bottom profiles are practically computed by using a training data set of size 800 text lines. Then the range of the value of the maximum number of black pixels from top and bottom profiles of all the three languages is obtained and used as a supporting feature. From the experimentation, it is observed that the value of topmaxrow for Kannada and English languages are overlapping and also the value of botmaxrow of Kannada and Hindi languages are overlapping. So there are chances of misclassifying an input text line when the values of the topmaxrow and botmaxrow of a text line lies in this overlapping range. This shows that only these two features are not enough to separate the three languages and hence it was intended to pick up two more features which give the location of the topmaxrow and botmaxrow. These two features are extracted by getting the row number of both top and bottom profiles holding the maximum number of black pixels. Thus the four features used for separating the three languages are –

**(i) Feature 1 - Topmaxrow:** The value of the range of maximum number of black pixels of top profile.

**(ii) Feature 2 - Topmaxrowno:** The row number of the top profile at which the maximum number of black pixels lies. **(iii) Feature 3 - Botmaxrow:** The value of the range of maximum number of black pixels of bottom profile and

**(iv) Feature 4 - Botmaxrowno:** The row number of the bottom profile at which the maximum number of black pixels lies.

The ranges of values of these four features are computed using a training data set of 800 text lines considering 300, 200 and 300 text lines from Kannada, Hindi and English languages respectively. The percentage of the presence of feature 1 - topmaxrow and feature 2 - botmaxrow are calculated. The range of values of all the four features obtained though experimentation are given in the Table I.

TABLE I. RANGE OF FOUR FEATURE VALUES (F1:FEATURE 1- TOPMAXROW; F2:FEATURE 2- TOPMAXROWNO; F3;FEATURE 3- BOTMAXROW; F4;FEATURE 4- BOTMAXROWNO).

| | Kannada | Hindi | English |
|---|---|---|---|
| **F1** | 38% to 55% | 58% to 80% | 34% to 40% |
| **F2** | 7 to 10 | 10 to 11 | 11 to 12 |
| **F3** | 24% to 30% | 25% to 32% | 33% to 42% |
| **F4** | 30 to 31 | 11 to 14 | 25 to 27 |

### *B. 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.

The different stages of the proposed method are:

Stage 1: Preprocessing: The input document image is segmented into several text lines and a bounding box is fixed by finding the leftmost, rightmost, topmost and bottommost black pixel of each text line. The bounded text line is resized to a standard size of 40X600 pixels.

Stage 2: Feature Extraction: The top-profile and the bottom-profile of each text line are obtained and the features are extracted from the profiles. The range of feature values obtained for the three languages Kannada, Hindi and English are computed and stored in a knowledge base.

Stage 3: Decision making: For a given test document image, the features are extracted and the values are computed. The values of the test images are compared with the values of the knowledge base and a rule based classifier is used to identify type of the language.

## III. EXPERIMENTAL RESULTS

The input document images used were downloaded from the internet. The size of the sample image considered was 512x512 pixels. The system is trained to learn the behavior of the top and bottom profiles with a training data set of 800 text lines, considering 300, 200 and 300 text lines from Kannada, Hindi and English languages respectively. We have tested our algorithm with a test data set of 600 text lines, having 200 text lines from each of the three languages. We have applied our algorithm on document images having both regular and italic font styles with different font sizes. Sample output images of Kannada, 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. From the Table-2, we can observe that the accuracy rate is high for Hindi text lines and the misclassification occurs between Kannada-English text lines and between Kannada-Hindi text lines. Figure II depicts the performance of recognition for a test data set of 800 text lines. A text line in Kannada is misclassified as English when more than 50% of the characters of that text line do not possess the head-line feature. 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. Through the experimental observation we have noticed that the misclassification amongst Kannada and English text lines is high when the font size of the text line is less than 16 and also when the size of the text line is less than 250X30 pixels. 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.

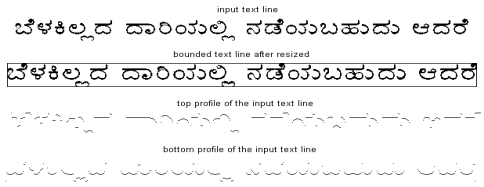


Figure 1. Sample output image of Kannada text line with its top and bottom profile.

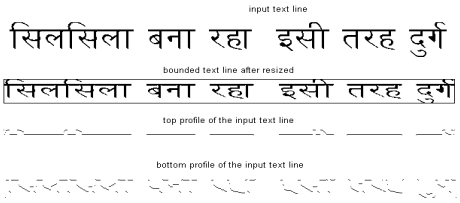


Figure 2. Sample output image of Hindi text line with its top and bottom profile.

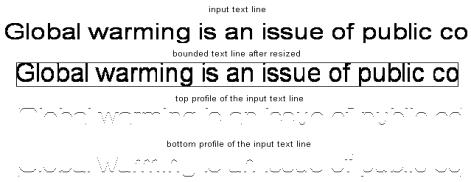


Figure 3. Sample output image of English text line with its top and bottom profile.

TABLE II. PERCENTAGE OF EXPERIMENTAL RESULTS.

| | **Kannada** | **Hindi** | **English** |
|---|---|---|---|
| **Kannada** | 93.7% | 3.5% | 2.8% |
| **Hindi** | 3.2% | 96.8% | 0% |
| **English** | 4.4% | 0% | 95.6% |

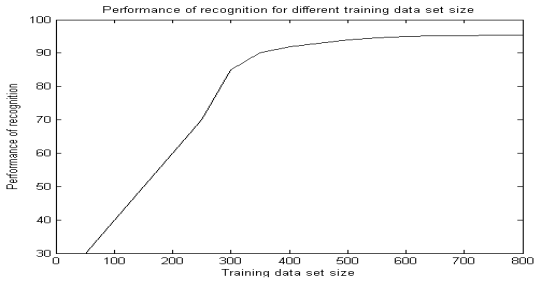


Figure 4. Sample output image of English text line with its top and bottom profile.

## IV. CONCLUSION

In this paper, an algorithm for language identification of Kannada, 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 Kannada, 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 Kannada, 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, "Identification and Separation of Text Words of Kannada, Telugu, Tamil, Hindi, English Languages through Visual Discriminating Features", in proc. of International conference on Advances in Computer Vision and Information Technology (ACVIT-2007), Aurangabad, India, 2007, pp. 1283-1291.