Volume : 2 | Issue : 6 | June 2013 • ISSN No 2277 - 8179

Research Paper

# Language identification of text words from tri-lingual document through Discriminating features

Engineering

**KEYWORDS :** Multi-lingual Document, Horizontal Lines, Vertical Lines, Feature Extraction.

**Priyanka P. Yeotikar** M.E. 2nd Year (I.T.), Sipna's COET, Amaravati, MS, India

**Prof. P. R. Deshmukh** Computer & I.T. Department, Sipna's COET, Amaravati, MS, India

## ABSTRACT

*For multilingual environment, multi lingual Optical Character Recognition system is needed to read multilingual documents. So, it is necessary to identify different language regions of the document before feeding the document to the OCRs of individual language. The objective of this paper is to propose visual clues based procedure to identify Kannada, Hindi and English text portions of the Indian multilingual document.*

## 1. INTRODUCTION

In India, single document page may contain words in two or more language scripts. So, multi-script OCR is necessary to read these documents for such a country. In most Indian script alphabet system apart from vowel and consonant characters, called basic characters, there are compound characters formed by combining two/more basic characters. The shape of compound character is usually more complex than the constituent basic characters. Many researchers have developed character recognizers tuned to specific applications, but multilingual capability has not received much attention. The capability of recognizing multilingual documents is both novel and useful.

One important task of document image analysis is automatic reading of text information from document image. The tool OCR performs this, which is broadly defined as process of reading optically scanned text by machine. Almost all existing works on OCR make an important implicit assumption that the script type of document to be processed is known beforehand. In automated multilingual environment, such document processing systems relying on OCR would clearly need human intervention to select the appropriate OCR package, which is certainly inefficient, undesirable and impractical. If document has multilingual segments, then both analysis and recognition problems become more severely challenging, as it requires the identification of the languages before analysis of the content could be made. So, pre-processor to OCR system is necessary to identify the script type of the document, so that specific OCR tool can be selected. The ability to reliably identify the script type using the least amount of textual data is essential when dealing with document pages that contain text words of different scripts. An automatic script identification scheme is useful to (i) sort document images, (ii) to select specific OCR systems and (iii) to search online archives of document image for those containing a particular script/language.

## 2. LITERATURE REVIEW

From the literature survey, it has been revealed that some amount of work has been carried out in script/language identification. Tan [2] has developed rotation invariant texture feature extraction method for automatic script identification for six languages: Chinese, Greek, English, Russian, Persian and Malayalam. In the context of Indian languages, some amount of research work on script/language identification has been reported. Santanu Choudhuri, [3] have proposed a method for identification of Indian languages by combining Gabour filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. Our survey for previous research work in the area of document script/language identification shows that much of them rely on script/languages followed by other countries and few from our country, but hardly few attempts focus on these three languages Kannada, Hindi and English followed in Karnataka, an Indian state. Keeping the drawback of the previous method in mind, we have proposed a system that would more accurately identify and separate different language portions of Kannada, Hindi and English documents and also to classify the portions of the document in other than these three languages into a fourth class category - OTHERS, as our intension is to identify only Kannada, Hindi and English. The system identifies the three languages in four stages: in the first stage Hindi is identified, in the second stage Kannada is identified, in the third stage English is identified.

## 3. IMPLEMENTATION STATEGY

### 3.1. Supportive Knowledge Base

A supportive knowledge base is constructed for each specific class of patterns, which further helps during decision making to arrive at a conclusion. The technique of obtaining the four visual features from the input image through experimentation is explained below:

1) Horizontal lines: In the binary image of each text line, if there are continuous one's in a row greater than the horizontal threshold value, then such continuous one's are retained resulting in a horizontal line and if there are no continuous one's greater than the horizontal threshold value, then such one's are changed to zeroes.
2) Vertical lines: In the binary image of each text line, if there are continuous one's in column greater than vertical threshold value then such continuous one's are retained resulting in a vertical line and if there are no continuous one's greater than vertical threshold value, then such one's are changed to zeroes.
3) Variable sized Blocks: The input binary image is segmented into several text lines and then each text line is segmented into several text words. Every text word is partitioned into three zones - upper zone, middle zone and lower zone to get upper line and lower line as two boundary lines for every text word. Then every text word is scanned vertically from upper line to reach the lower line of the respective text word without touching any black pixel, which results in a stream of variable sized blocks.
4) Blocks with more than one component: The number of components present within each block is computed using 8-neighbour connectivity. The percentage of spatial occurrence of all the four visual features for each of three languages are practically computed through extensive experimentation and stored in the knowledge base.

### 3.2. Line-wise Identification model

We have proposed line level identification model that would accurately identify and separate different language portions of Kannada, Hindi and English text lines of the input document and also group the portions of the document in other than these three languages into a separate class called OTHERS, without identifying the type of the language, as our intention is to identify texts in Kannada, Hindi and English languages only.

The proposed model is developed based on the discriminating features viz., horizontal lines, vertical lines, variable sized blocks and blocks with more than one component. These discriminating features are extracted from the processed document image and compared with the values that are stored in the knowledge base, to arrive at a decision regarding the type of the text language.

Volume : 2 | Issue : 6 | June 2013 • ISSN No 2277 - 8179

The different steps involved in implemented model are as follows:

Input: 256x256 JPEG file containing text lines in Kannada / Hindi / English

Output: Text line of Kannada, Hindi and English languages.

Step-1: The input document is preprocessed noise removed, smoothing done, skew compensated and binarized.

Step-2: Line segmentation: To segment document image into several text lines, we use the valleys of the horizontal projection computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes one boundary line. Using these boundary lines, document image is segmented into several text lines.

Step-3: Zonalization: Each text line is partitioned into three zones - upper zone, middle zone and lower zone .

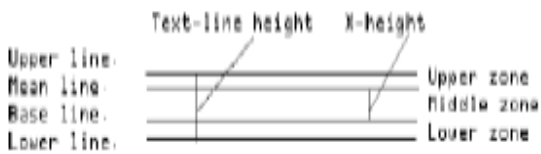


**Figure.3.1 Text Line partitioned into three zone**

Step-4: Block segmentation: From zonalized text line, upper line and lower line is used as two boundary lines for every text line. Then every text line is scanned vertically from its upper line to reach its lower line without touching any black pixels to get a boundary line. Such characters enclosed within each boundary lines lead to a stream of blocks.

**Step-5: Feature extraction:**

(i): Horizontal line detection: From the input image, the horizontal lines are obtained. Then the percentage of the presence of these horizontal lines for each text line is computed and compared with stored values in the knowledge base.

(ii): Vertical line detection: From the input image, the vertical lines are obtained Then the percentage of the presence of these vertical lines for each text line is computed and compared with the stored values in the knowledge base.

(iii): Variable Sized blocks: The size of blocks of each text line is calculated by taking the ratio of width to height of each block. Then the percentage of equal and unequal sized blocks of each text line is calculated.

(iv): Blocks with more than one component: The percentage of the number of components present in each block of every text line is computed.

**Step-6: Decision making:**

(i) Condition-1: If 90% of horizontal lines on the mean line is greater than two times the X-height of the corresponding text line; if there are 80% of vertical lines in the middle zone and also if 70% of the blocks have width greater than two times the X-height, then such portion of document is recognized as Hindi language.

(ii) Condition-2: If 65% of horizontal lines on the mean line is greater than half of the X-height of the corresponding text line and if there are 40% of unequal sized blocks in a text line, then such portion of document is recognized as Kannada language.

(iii) Condition-3: If there are 80% of vertical lines in the middle zone greater than half of the text line height and if 80% of the blocks are equal in size, then such portion of document is recognized as English language.

(iv) Condition-4: If output image does not belong to any of the above three classes, then such portion of the document is grouped into a separate class called OTHERS.

## 4. RESULTS

The implemented model has been tested on test data set of 400 document images containing about 600 text words from each script. The proposed model is tested for various font types and the results are given in Table 4.1

| Script type | Font style | No. Of Samples | Correct Recognition | Recognition Rate |
|---|---|---|---|---|
| Kannada | Siriganda | 150 | 145 | 96.66% |
| | Kasturi | 150 | 148 | 98.66% |
| | Vijaya | 150 | 147 | 98% |
| Hindi | Vijaya | 200 | 198 | 99% |
| English | Times new Roman | 100 | 98 | 98% |
| | Arial | 100 | 98 | 98% |
| | Bookman Old style | 100 | 98 | 98% |

**Table 4.1**

The success rate of percentage of recognition of all the three scripts is given in Table 4.2 from the experimentations on the test data set; the overall accuracy of the system has turned out 98.7%. The performance of the proposed model is evaluated from scanned document images also. The overall accuracy of the system reduces to 98.4% due to noise and skew-error in the scanned document images.

| Script Types | Manually Created Data set | | | Printed Scanned Data set | | |
|---|---|---|---|---|---|---|
| | Classified | Misclassified | Rejected | Classified | Misclassified | Rejected |
| Kannada | 97.7% | 0.8% | 1.5% | 95.5% | 2.5% | 3% |
| Hindi | 99% | 0.5% | 0.5% | 97% | 1.6% | 1.4% |
| English | 98% | 0.5% | 0.7% | 97.6% | 0.8% | 1.6% |

**Table 4.2**

## 5. CONCLUSION

In the proposed system we have implemented line-wise identification model to identify Kannada, Hindi and English text words from Indian multilingual machine printed documents by using to-down & bottom up approach. The proposed models are developed based on four visual discriminating features, which serve as useful visual clues for language identification. The performance of proposed algorithm is encouraging when the algorithm is tested using manually created data set. The system exhibits an overall accuracy of 98.4%. The work could be extended to character level script identification and for other Indian scripts.

## REFERENCE

1]. P.Naghabhushan, Radhika M Pai, "Modified Region Decomposition Method and Optimal Depth Decision Tree in Recognition of non-uniform sized characters – An Experimentation with Kannada Characters", Journal of Pattern Recognition Letters, (1999). | [2]. T.N.Tan, "Rotation Invariant Texture Features and their use in Automatic Script Identification", IEEE Trans. Pattern Analysis and Machine Intelligence (1998). | [3]. Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet, "Identification of Scripts of Indian Languages by Combining Trainable Classifiers", Bangalore, India. | [4]. M.C.Padma, P.Nagabhushan, "Horizontal and Vertical linear edge features as useful clues in discrimination of multiligual machine printed documents", National Workshop on Image Processing , Madhurai, (2002). | [5]. U.Pal, B.B.Choudhuri, "OCR in Bangla:an Indo- Bangladeshi language", IEEE, no.2, 1051-4651, (1994). |