Lateral Cross-sectional Analysis based Classification of Bi-lingual Malayalam-English OCR using Distinct Features and non- Uniform Sub-classes

Bindu Philip
JSS Research Foundation
S J College of Engineering
Mysore - 570006
Karnataka, INDIA
+91-821-2511383 e2323
binduthomas25@yahoo.co.in

R. D. Sudhaker Samuel
Department of Electronics & Communication
S J College of Engineering
Mysore - 570006
+91-821-2511383 e2212
sudhakersamuel@yahoo.com

ABSTRACT

Feature extraction, to reduce dimensionality is a crucial process in the development of any robust optical character recognition system. In India bilingual documentation is very common especially government forms and formats, technical documents, postal documents, railways reservation forms etc. are bilingual and at times trilingual. Even documents printed in a single language often contain English words and numerals. An integrated recognition system that recognizes characters as well as numerals belonging to different languages in a single document has innumerous emerging applications and hence the motivation to work further in this area. Indian languages especially South Indian languages have several distinct characteristics which could be exploited to define features. In this paper Malayalam language is chosen as its script has exceedingly rich features. The length of the characters and the occurrence of transverse strokes in lateral directions turn out to be promising distinct features of these characters. This paper presents a lateral cross sectional analysis based approach for the recognition of Bi-lingual characters. Analysis is performed keeping track of the number of impulses at edges along each row in the image matrix resulting in distinct features. These features are grouped into subclasses based on heuristics depending on intra-subclass distances thus creating sub-classes of varying lengths. The feature selection method reduces computational complexity without compromising efficiency. A simple L_2 – distance based classifier gives good performance.

Categories and Subject Descriptors

I.7.5 [**Document and text processing**]: Document Capture – *Document analysis, Graphics recognition and interpretation*, *Optical character recognition (OCR), Scanning.*

General Terms

Algorithms, Performance, Design, Reliability, Experimentation.

Keywords

Bi-lingual optical character recognition, Lateral cross-sectional analysis, Intra-subclass distances, Distinct features, Projection profile method.

1 INTRODUCTION

OCR technology gives scanning and imaging systems the ability to convert images of machine-printed characters into machine-readable characters. OCR systems aim at enabling machines to recognize optical symbols without human intervention [1]. Character extraction and recognition techniques have potential application in any domain where massive document image-bearing texts must be interpreted, analyzed and processed [12]. Postal automation is one such application and is vital to render postal services more efficient. The multilingual nature of our country poses an additional challenge [13]. In literature, most of the work is seen to be carried out for detection of English text from images using classical approaches, which have limitations when applied to Indian languages. There are 17 officially recognized languages in India: Hindi, Marathi, Sanskrit, Punjabi, Gujarati, Oriya, Bengali, Assamese, Telugu, Kannada, Malayalam, Tamil, Konkani, Manipuri, Urdu, Sindhi and Kashmiri. Most languages are written in their own script. Indian scripts are rich in patterns and variations. At the same time these variations often point to the solution. On keen observation it is noticed that the transverse strokes caused by the curly and curvy structure of these characters can be used as a distinct feature for classification. In this paper Malayalam was preferred as it has remarkably distinct lateral variations as compared to many other Indian languages. This attribute provides the basis for the selection of the most favorable feature vectors for classification of these characters. Added to this Malayalam script characters have varying number of columns when converted to its bitmap ranging from 53 columns to a phenomenal 346. This unique characteristic of large variations in the number of columns is used for coarse classification.

The character with smallest number of columns, *ma*, is of size 53 x 61 (Figure 1),

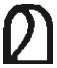

Figure 1: Character with smallest number of columns

while the character with largest number of columns, *jhau*, is of size is 347 x 67 (Figure 2).

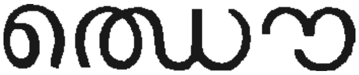

Figure 2: Character with largest number of columns

Selection of distinct features, its extraction and classification play a pivotal role. A standard database of features of all the 592 extended Malayalam characters, 62 uppercase, lowercase English characters and numerals, totaling 654 is created. A training set is formed and the feature vector of each is extracted based on lateral cross-sectional analysis to be explained in subsequent sections. The simplest of the classifier, based on L_2 distances is used to recognize the characters at the moment.

2 THE MALAYALAM SCRIPT

Malayalam script is derived from the Grantha script, a descendant of the ancient Brahami script and is the language of Kerala state. It ranks eighth in the list of 15 most popular languages in India. Malayalam alphabets are classified into two categories: vowels and consonants. Some notable features of this language are it has a syllabic alphabet set in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel [10]. When they appear at the beginning of a syllable, vowels are written as independent letters. When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter. Examples are shown in Figure 3.

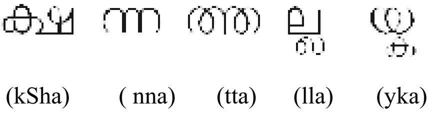

Figure 3: A few conjunct characters

Malayalam is written horizontally from left to right and its basic set of symbols consists of 36 consonants and 14 vowels. The composite characters and extended characters put together make a total of 592 characters in Malayalam.

3 FEATURE EXTRACTION

Classical feature extraction methods determine an appropriate subspace of dimensionality in the original feature space. Linear transforms such as principal component analysis [7][14], linear discriminate analysis [14], factor analysis and projection pursuit have been popularly used in pattern recognition for feature extraction and dimensionality reduction. The text image having a combination of Malayalam and English characters along with English numerals is scanned and subjected to preprocessing where the unwanted pixels are removed. Further, segmentation is done based on the classical projection profile procedure to extract lines, words and characters (Malayalam and English) and numerals. These segmented characters are then normalized to a height of 50 pixels preserving the length of the characters. There are several efficient normalization methods [4][5][6][9]. The most interesting stage of recognition is the extraction of distinct lateral features. The components of the resultant feature vector are obtained by keeping track of the intensity transitions along each row. The preprocessed, segmented and normalized character is now represented by a matrix A,

Where $A = \left(a_{ij}\right) \in R^{mxn} : a_{ij} = \{0,1\}$ (1)

For practicality in classification and identification, reduction in dimensionality is performed in order to obtain a feature vector mxn $x \in R^m$, at the same time capturing the distinct information in all the '*n*' columns. Typical feature extraction methods such as Hu moments [8], LDA [11][14] and PCA[7][14 work on a normalized mxm image matrix where vital lateral information is lost. The work presented in this paper captures useful information along the rows and hence is called lateral analysis. The key issue here is to retain selected features along rows instead of losing them by compression or by looking only at overall characteristic feature of a matrix. The matrix 'A' as represented in equation (1) has several lateral features.

One such feature has been efficiently exploited in the work reported in this paper. It essentially captures the frequency of transitions along each row, to form the feature vector, $x \in R^m$ of the matrix $A \in R^{mxn}$, given by

$$x_i = \sum_{j=1}^{n} | a_{i,j+1} - a_{i,j} | \qquad (2)$$

Thus corresponding to the 654 characters, there would be 654 image feature vectors given by

$$\left(x^k\right) \in R^m \quad \text{Where, k=1,2,…..654}$$

The uniqueness of these k feature vectors have been ascertained by arranging the computed values of $|| x^k || \forall k$ in descending order. It was seen in the list that between subsequent entries $||| x^k || \sim || x^{k-1} ||| \geq \xi \ \forall \ k$ ensuring that these features are useful for classification, where $x^k, x^{k-1}, ...$ etc now are feature vectors in reducing order of magnitudes of $|| x ||$. Further, it was observed that ξ could be improved by arranging the characters in the order of their number of columns *'n'*. The corresponding features are grouped into 12 subclasses based on heuristics depending on intra-subclass distances thus creating windows of varying lengths. A database of features of all 592 extended Malayalam characters, 62 uppercase, lowercase English characters and numerals are created.

4 THE L2-CLASSIFIER

The objective of classification is to build a set of models that can correctly locate the class of different objects. In methods like the k-nearest-neighbor classification, the input is a set of objects (training data), the classes to which these objects belong to (dependent variables) and a set of variables describing different characteristics of the objects (independent variables). Once such predictive model is built, which is used to predict the class of the objects for which class information is not known a priori. In our implementation even a simple Lp – distance based classifier seems to give good performance. Classification is performed in two stages which further speeds up the rate of recognition. In the first stage, the number of columns of the test image is used to locate the closest subclass of the test characters (coarse classification) while the Lp–based classifier (fine classification) uses the test feature vector extracted by applying the lateral analysis to locate the nearest vector in the subclass in order to classify the test character.

5 IMPLEMENTATION MODEL

The block diagram of the Bilingual character and numeral recognition engine is shown in Figure 4.

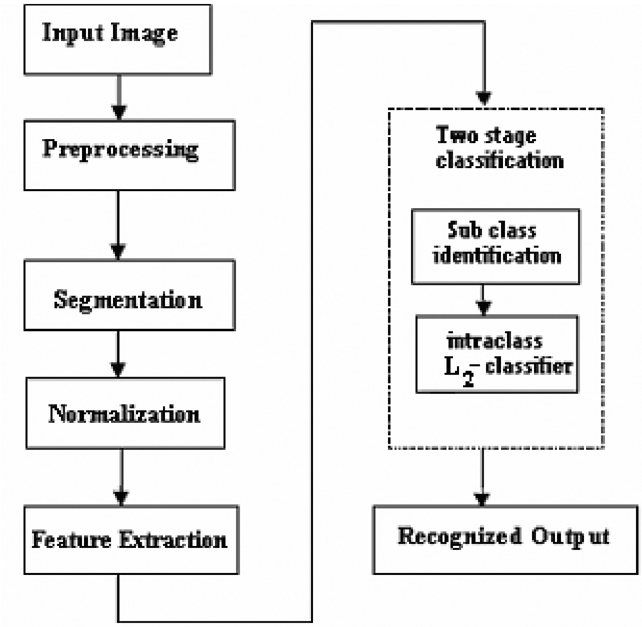

Figure 4: Block diagram of the proposed Bi-lingual recognition system

The stages involved in the development of the model are image acquisition, preprocessing, segmentation, normalization, feature extraction and classification. A printed document containing a blend of complete composite Malayalam, English characters and numerals is scanned on a flatbed scanner (600 dpi) for digitization. This digitized image is preprocessed for removal of background noise and a bitmap image is obtained. Malayalam is a non-cursive script where the individual characters in a word are naturally isolated. Classical horizontal and vertical profile method [14] is used for segmentation which involves, line segmentation followed by word segmentation and later character or number segmentation. The segmented fragments are now normalized to a height of '*m1*' pixels preserving the length of the characters. There are several efficient normalization methods [3][4]. It is significant to mention that '*m1 = 50*' was found to be an optimum value for the moment considering efficiency and complexity as performance parameters. The problem now reduces to the characterization of mxn image matrices. The value of '*m*' however can be fixed at an optimal value to obtain distinct features of the entire data set. The classical methods of characterization include LDA [11][14], PCA[7][14], Hu moments[8], SVD etc. These were successfully implemented with some modifications [2][3]. A two-stage classification is performed to accelerate the rate of recognition and reduce complexity to a great extent. Heuristics is again applied to remove ambiguities in feature vectors lying on the boundaries of the sub classes.

6 RESULTS AND ANALYSIS

Feature vectors of all the 592 extended Malayalam characters, 62 uppercase, lowercase English characters and numerals forms the training data set. A notable outcome of lateral analysis is that it is inherently invariant to font face and proper normalization ensures invariance to font size. The paper has addressed the problem of structure- based Bi-lingual character recognition. Recognition of Malayalam, English characters and numerals based on distinct features has been demonstrated with good accuracy. The approach has been tested successfully on all the 654 elements of the data set with 4 to 5 test images of each character or number. Right classification has been obtained in all the trials. The feature vectors of composite Malayalam character set, uppercase and lowercase English characters and the numerals are grouped into 12 subclasses based on the number of columns of the character or numeral in the data set.

Analysis is made on characters with close resemblance and falling into the same subclass. Table 1 shows that the L_2 distances are large between these identical appearing characters.

Table 1: L_2 Distances between similar characters

Sl No	Characters		L_2 Distance
1.	ഋ (RU)	ൠ (Ru)	11.9164 Units
2.	എ (e)	ഏ (ee)	3.87298 Units
3.	ഡ (da)	ഢ (edha)	12.2882 Units
4.	ടീ (tee)	ടി (ti)	8.7178 Units
5.	ചം (cham)	ചഃ (chaH)	16.7033 Units
6.	ടു (tu)	ടൂ (too)	5.38516 Units
7.	ചം (cham)	പം (pam)	7.7459 Units
8.	റീ (rxee)	റി (rxi)	10.0499 Units
9.	രൈ (rai)	റൈ (rxai)	11.8743 Units
10.	റി (rxi)	രി (ri)	11.5758 Units

A few characters or numeral with the same number of columns within the subclass are also considered for analysis. The L_2 distances between these elements shown in Table 2 are found to exhibit remarkable difference and hence a notable improvement in the efficiency of the recognizer.

Table 2: L_2 Distances between characters with same number of columns and within the subclass

No. of cols	Characters		L_2 Distance
188	ഷെ (She)	ബേ (bE)	18.0 Units
188	ബേ (bE)	ബീ (bI)	27.3496 Units
188	ബീ (bI)	ബൃ (bRu)	32.0936 Units
188	ബൃ(bRu)	ഘൂ (GU)	19.1372 Units
188	ഘൂ (GU)	ഷെ (She)	50.448 Units
233	ഫോ(PO)	ഞോ(GO)	19.0788 Units
233	ഞോ(G0)	ജൈ(jai)	10.0499 Units
233	ഭൌ(Bau	ഞോ(GO)	20.9762 Units
233	ഭൌ(Bau	ഷോ(Sho)	18.6011 Units

The histogram of column widths of the 654 characters is shown in Figure 5.

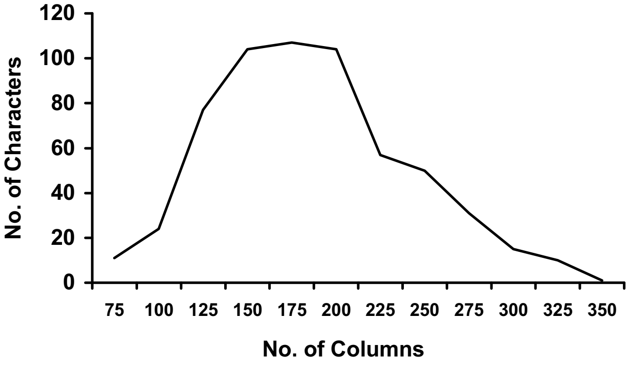

Figure 5: Histogram of column widths of the 654 characters

This skewed Gaussian distribution of column widths is used to select the subclass boundaries.

We assertively conclude that lateral analysis is a good option for classification of characters in languages with characters over a wide column width with distinct lateral features.

7 ACKNOWLEDGEMENTS

This work was supported in part by research grants from UGC for Major Research Project in Science and Technology, F.No. 32-113/2006. Our thanks to "[Dr R Vittal Rao, Indian Institute of Science, Bangalore, India, Personal Communication]".

8 REFERENCES

[1] K.H.Aparna, Sumanth Jaganathan, P.Krishnan, V.S.Chakravarthy, An optical Character Recognition System for Tamil Newsprint, International Conference on Universal Knowledge and Language 2002

[2] Bindu Philip and R D Sudhaker Samuel, "A Malayalam Character Recognition System based on Feature Subset Selection Method using Principle Component Analysis and Linear Discriminate Analysis with k nearest neighbor classifier", National Seminar On Intelligent Communication And Intelligent Systems.

[3] Bindu Philip and R D Sudhaker Samuel, "A Hu's Moment based K-NN Classifier for a Novel and Fast Bi-lingual Character Recognition System for Malayalam and English Languages", in the proceedings of the National Workshop on Intelligent Data Analytics & Image Processing.

[4] Christophe Choisy and Abdel Belaid, Handwriting Recognition Using Local Methods for Normalization and Global Methods for Recognition, *Proceedings of the IEEE*, 2001, 23- 27.

[5] Cheng-Min Cho and Shuenn-Shyang Wang, A New Fuzzy Normalization Algorithm for Handwritten Chinese Characters Recognition, *Proceedings of the IEEE, 2002,* 368-371

[6] Geetha Srikantan, Dar-Shyang Lee and John T. Favata, Comparison of Normalization Methods for Character Recognition, *Proceedings of the IEEE*, 1995, 719-722.

[7] Myoung Soo Park, Jin Hee Na and Jin Young Choi, PCA-based Feature Extraction using Class Information, Proceedings of the IEEE, 2000, 883- 887.

[8] M. K. Hu, *pattern recognition by moment invariants*, Proc. IEEE, 1961, pp.1428

[9] Jose Joesmar de Oliverira Jr.,Luciana R. Velso and Joao M. de Carvalho, Interpolation / Decimation scheme applied to size Normalization of characters Images, *Proceedings of the IEEE*, 2000,577-580.

[10] P. S. Janardhanan. Issues in the development of OCR systems for Dravidian languages - proceedings of Akshara 94., BPB Publications, New Delhi, India 1994.

[11] Jian Yang David Zhang and Zhong Jin Jing-yu Yang, Unsupervised Discriminant Projection Analysis for Feature Extraction, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), 2006.

[12] Mori, *et al.* 'Optical Character Recognition: A Survey.' *Proceedings of the IEEE*, 1992.

[13] U. Pal and B. B. Chaudhuri, Indian Script Character Recognition: A Survey Pattern Recognition, 2004, 1887-1899

[14] O. D. Trier, A. K. Jain and T. Taxt, "Feature Extraction Methods for Character Recognition: A Survey", Pattern Recognition, vol. 29, no. 4, 1996, pp. 641-662