

An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers

Bindu Philip¹ and R. D. Sudhaker Samuel²

^{1,2}Department of Electronics & Communication, S J College of Engineering, Mysore, India
Email: ¹binduthomas25@yahoo.co.in and ²sudhakersamuel@yahoo.co m

Abstract—This paper describes an Optical Character Recognition (OCR) System for printed text documents in Malayalam, a South Indian language. Indian scripts are rich in patterns while the combinations of such patterns makes the problem even more complex and these complex patterns are exploited to arrive at the solution. The system segments the scanned document image into text lines, words and further characters and sub-characters. The segmentation algorithm proposed is motivated by the structure of the script. A novel set of features, computationally simple to extract are proposed. The approaches used here are based on the distinctive structural features of machine-printed text lines in these scripts. A lateral cross-sectional analysis is performed along each row of the normalized binary image matrix resulting in distinct features. The final recognition is achieved through classifiers based on the Support Vector Machine (SVM) method. The proposed algorithms have been tested on a variety of printed Malayalam characters and currently achieve recognition rates between 90.22% and 95.31 %.

Index Terms—Malayalam Script, OCR, Structural approach, Segmentation, Support Vector Machine (SVM) Classifier.

I. INTRODUCTION

Optical character recognition system has received considerable attention in recent years due the tremendous need for digitization of printed documents. In this paper we describe a document image analysis system that can handle printed text documents in Malayalam, the official language of the south Indian state of Kerala. The input to the system is the scanned image of a page of printed Malayalam text. The output is an editable computer file containing the text data in the printed page. The most important phases involved are Segmentation and Feature Extraction. The task of separating lines and words in the document is fairly independent of the script and hence can be achieved with conventional projection profiles techniques. However, due to the peculiarities of the Malayalam script, a novel segmentation scheme is proposed whereby words are first segmented to a sub-character level and the individual pieces are recognized. These are then put together for the recognition of the characters. The proposed system employs a classifier based on the concept of Support Vector Machines (SVM).

Currently there are many OCR systems available for handling printed English documents with reasonable levels of accuracy. (Such systems are also available for

many European languages as well as some of the Asian languages such as Japanese, Chinese etc.) However, there are not many reported efforts at developing OCR systems for Indian languages.

An automatic character recognition system is one of the most fascinating and challenging areas of pattern recognition with a wide range of practical applications like mail sorting, forms processing, preserving historical documents in editable format, desktop publication, backup files of rare books, reading aid for blind, and other applications involve language processing, word indexing, library automation. The different challenges that exist in Malayalam script are its large character set of roughly more than 900 characters, similarity of character shapes, and complexity of character structure. This paper addresses all these challenges.

II. SOME EXISTING OCR TECHNIQUES FOR INDIAN SCRIPTS

Some of the existing techniques used in OCR for Indian scripts work is presented here. Pal & Chaudhuri [2] reported a complete OCR system for printed Devnagari here headline deletion is used to segment the characters from the word. An OCR for Telugu is reported by Negi, et. al[4], where instead of segmenting the words into characters as usually done, words are split into connected components (glyphs). Some contributions that report the use of SVM classifier are, a font and size independent OCR system for printed Kannada documents using support vector machines reported by Ashwin T V and P.S Sastry [1]; Seethalakshmi, et. al, [3], reported a Tamil OCR using Unicode and SVM classifier. Renju John, et.al., [6] reported work on isolated Handwritten Malayalam Character Recognition based on 1 D Wavelet Transform. Recognition of Isolated handwritten character images based on k-nearest neighbour classifier is reported by Lajish, et.al., [5]. A comprehensive study on the success rate of well known feature extraction methods in terms of recognition accuracy and computational complexity is yet to be reported. There is hardly any reporting on techniques used for printed Malayalam OCR.

III. MALAYALAM SCRIPT

Malayalam is a Dravidian language with about 35 million speakers. It is spoken mainly in the south western

India, particularly in Kerala. The Malayalam script is derived from the Grantha script, a descendant of the ancient Brahmi script. The character set consists of 13 vowels, 2 left vowel signs, 7 right vowel signs, some appear on both sides of the Conj/consonant, 30 commonly used conjuncts, 36 consonants and vowel signs are shown in Figure 1. The dependent vowels do not stand on their own, but are depicted in combination with a consonant or consonant cluster [7]. The positioning of the dependent vowel may be to the left, to the right, or both to the left and right of the consonant/conjunct, depending on the vowel sign being attached [8] as shown in Figure 1.

Malayalam has remarkably distinct lateral variations as compared to many other Indian languages with a number of curls and twists in the characters. Another very

vowel sign	left/right of the consonant/conjunct	(vowel sign attached to ക) example
ഓ	right	കഓ
ഐ	right	കഐ
ഓ	right	കഓ
ഐ	right	കഐ
ഓ	right	കഓ
ഐ	right	കഐ
ഐ	left	ഐക
ഓ	left	ഓക
ഐ	left	ഐക
ഐ ഓ	left and right	ഐകഓ
ഓ ഓ	left and right	ഓകഓ
ഐ	right	കഐ

Figure 1: Malayalam Vowel Signs

interesting feature of this script is that the number of columns varies from 53 to a phenomenal 347 columns over the entire extended character set of the language.

IV. IMPLEMENTATION MODEL

The stages involved in the development of the OCR engine are image acquisition, preprocessing, segmentation, normalization, feature extraction and classification. A printed document containing Malayalam text is scanned on a flatbed scanner at 300 dpi for digitization. This digitized image is preprocessed for removal of background noise and the grey scale image is converted to a binary image after which line segmentation and word segmentation is performed using classical horizontal and vertical projection profiling technique. Character segmentation is then done using a novel segmentation algorithm as explained in Section V. The characters and sub characters in printed Malayalam text have uniform distance of separation and thus segmentation of full characters in a Malayalam word is a great challenge. The segmented fragments are now normalized to a height of m_1 pixels preserving the length

of the characters. It is significant to mention that $m_1 = 50$ was found to be an optimum value after several trials. The problem now reduces to the characterization of $m \times n$ image matrices. The value of m however can be fixed at an optimal value to obtain distinct features of the entire data set economically. The character recognition engine now performs feature extraction by finding a set of vectors, which effectively represent the information content of a character. A novel set of features, computationally simple to extract are proposed and the printed Malayalam characters are classified using hierarchical SVM classifiers.

V. SEGMENTATION

A Malayalam character could consist of several uniformly spaced unconnected components such as a vowel (only at the beginning of a word) or a consonant/conjunct along with vowel signs. Conventional techniques like horizontal and vertical projection profile methods fail to segment the complete character correctly because of the equal space between the characters of a word and the sub characters of a character within a word. An appropriate and novel technique is proposed in this paper to segment Malayalam Characters. We considered the fact that printed Malayalam characters can have a maximum of three segments. The first segment could have either none or possibly one or two left vowel signs (a unique case). The second segment would be the core character which could be either a vowel or a consonant or a conjunct while the third segment could again have either none or one of the seven right vowel signs as shown in Figure 1. An example of a character with all three segments is shown in Figure 2.

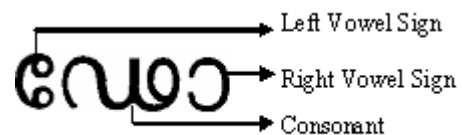


Figure 2: A Typical Malayalam Character

Let X represent a consonant/ conjunct and Y represent a vowel. Further let 0 represent vowel signs appearing to the left of the consonant/ conjunct while 1 represent vowel signs appearing to the right of the consonant/conjunct. The valid character sequences are of the form Y, X, X1, 0X, 0X1 and 00X, where each of these sub characters Y, 0, X, and 1 are segmented out using the classical vertical projection profile method. These segmented sub characters are applied to the logic shown in the flow chart of Figure 3 and the classification search space as shown in Figure 4 where V represents vowels(13 in all), C represents consonants (36 in all), Conj represents conjunct characters(30 in all), VS_L represents the vowel signs to the left(2 in all) while VS_R represents the vowel signs to the right(7 in all). The end of the word is identified by a larger valley in the vertical projection profile. Note that the sub characters are extracted using the smaller valley of projections. All the sub characters of a word are subjected to this logic in

sequence of appearance in the word. The logic used to reduced the search subspace is that the first level search subspace has vowels, consonants, left vowel signs and conjuncts. The first recognized character/ sub character of a word falls into one of these four categories only.

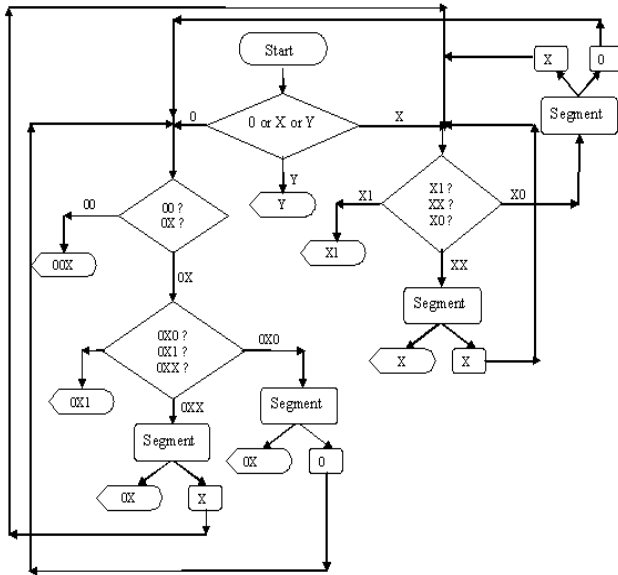


Figure 3: Segmentation Logic Flow Chart

The search space further reduces for finding the subsequent character or sub character as independent vowels can appear only in the beginning of a word. The logic used behind the choice of search space for the classifier is based on the sequence of arrangement of the segments of the character. This logic facilitates accurate segmentation.

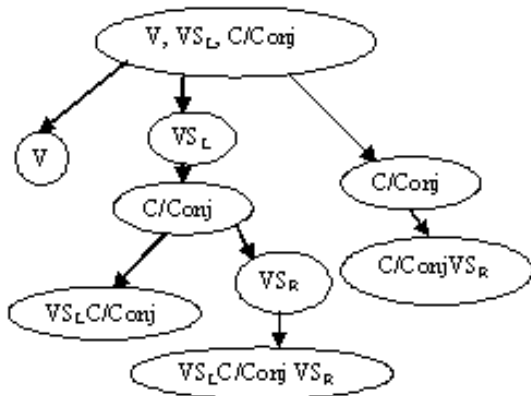


Figure 4: Classification Search Space

VI. FEATURE EXTRACTION

The process of digitization followed by segmentation essentially renders the image in the form of an $m \times n$ matrix. These matrices are then generally normalized and then converted into a square matrix in order to apply the classical tools of linear algebra for characterization. It is easy to see that any form of characterization results in a reduction in dimensionality which essentially helps in the search process by classification over a large data base.

However, there are instances where rich information along rows would be lost in the process of reducing the image matrix to square.

One good example is the segmented images of the characters of Malayalam language. It thus makes good sense to retain the number of columns. A rectangular, strictly black and white digital image preprocessed to remove any extraneous noise can be represented by a matrix A, where

$$A = (a_{ij}) \in \mathbb{R}^{m \times n} : a_{ij} = \{0,1\}, \quad (1)$$

usually $n > m$.

In order to ensure practicality in classification and identification, for the matrix in equation (1) reduction in dimensionality is performed to obtain a feature vector $x \in \mathbb{R}^m$, at the same time capturing the distinct information in all the n columns. The methods proposed in this paper essentially capture useful information along rows of matrixes and hence we call them lateral analysis. Selected features along rows are retained rather than losing them by compression, normalization or by looking only at the overall characteristic feature of a matrix. The matrix A as represented in equation (1) has several lateral features. Three of these are presented here.

A. Frequency Capture

This process in principle captures the frequency of transitions along each row. The feature vector, $x \in \mathbb{R}^m$ of the matrix $A \in \mathbb{R}^{m \times n}$ in this approach is defined by

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

This captures features of characters with multiple loops which is a distinct feature of Malayalam characters.

B. Average Gap Analysis

Here, the average gap along each row is computed to get feature vector $x \in \mathbb{R}^m$ of the matrix $A \in \mathbb{R}^{m \times n}$ given by

$$x_i = \frac{n - \sum_{j=1}^n a_{i,j}}{\sigma_i} \quad (3)$$

$\sigma_i = (\sum_{j=1}^{n+1} |b_{i,j+1} - b_{i,j}|) / 2$ with each row padded

with 1s on either ends for the sake for computational convenience, where

$$(b_{ij}) = (a_{ij}) \text{ for } i=1\dots m \text{ and } j=2\dots n+1$$

$$(b_{i,1}) = (b_{i,n+2}) = 1 \forall i$$

This captures gaps between the numerous curls in the characters, once again a distinct characteristic of Malayalam characters.

C. Absorption

Here, the number of ones along each row is computed to get feature vector $x \in \mathbb{R}^m$ of the matrix

$$A \in \mathbb{R}^{m \times n} \text{ given by } x_i = \frac{\sum_{j=1}^n a_{i,j}}{n} \quad (4)$$

This essentially captures the large number of vertical strokes, typical of Malayalam characters.

The features extracted using these methods are grouped into sub classes and provided to the classification module.

Let the data base consist of k images. These k images would correspond to k image matrices as represented by equation (1). Using any of the three approaches, k feature vectors $(x^k) \in \mathbb{R}^m$ are computed. It was seen that for consecutive vectors $\| \|x^k\| - \|x^{k-1}\| \| \geq \xi \forall k$ ensuring that these features are sufficiently distinct to be useful for classification by using SVM classifier. The method was successfully applied to the OCR of the complete Malayalam character set. Malayalam has exceedingly rich patterns along rows. The transverse stroke caused by the curly Malayalam characters attracts it to the proposed methods. Further for a given font size the Malayalam character range from 53 columns to a phenomenal 346.

VII. CLASSIFICATION

The Support Vector Machine (SVM) classifier in its basic form implements two-class classifications. The objective is to further improve the recognition rate by using support vector machine (SVM) at the segment classification level. The advantage of SVM, is that it takes into account both experimental data and structural behavior for better generalization capability based on the principle of structural risk minimization (SRM)[9]. Its formulation approximates SRM principle by maximizing the margin of class separation, the reason for it to be known also as large margin classifier.

The fundamental idea of SVM classifiers is to find a separating hyperplane between two classes ($h : w^T x + b = 0$), so that minimal distance with respect to the training vectors, called margins, is maximum.

The optimal solution is obtained when this hyperplane is located in the middle of the distance between the convex envelopes of the two classes. This distance is denoted by d_m and is expressed by,

$$d_m = \frac{2}{\|w\|}$$

The support vectors are situated on the margins of the two classes.

If the training vectors membership is defined by

$$u_k = 1 \quad \text{if } x_k \in \omega_1$$

$$u_k = -1 \quad \text{if } x_k \in \omega_2$$

Then the support vectors can be written in the form

$$\Omega_s = \{x_k \mid u_k (w^T x_k + b) = 1\}$$

The structure of the SVM classifiers can be modified to also generate non-linear separating surfaces. The basic idea is to project the input vectors in higher dimension space where the classes become linearly separable. This transformation is performed by means of a non-linear function Φ with modifies the scalar products of the two input space vectors.

$$x_k \rightarrow \Phi(x_k) \text{ and } x_j \rightarrow \Phi(x_j)$$

$$\Rightarrow x_k^T x_j \rightarrow \Phi(x_k)^T \Phi(x_j)$$

the function Φ is replaced by a symmetric and separable function called kernel Δ .

The Kernel Function is defined as

$$\Delta(x_k, x_j) = \exp(-\alpha \|x_k - x_j\|^2)$$

$$\Delta : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \Delta(x_k, x_j) = \Phi(x_k)^T \Phi(x_j)$$

The performance of SVM depends on the kernel. We use RBF (Gaussian) kernel, which out performed the other commonly used kernels in the preliminary experiments. Gaussian RBF kernel is given as

$$\Delta(x_k, x_j) = \exp(-\alpha \|x_k - x_j\|^2)$$

A. Implementation of SVM Classifier

The classification stage is the decision step and associates a label with an input pattern. The segmentation of the characters into segments the word into vowels, consonants, conjuncts, left vowel signs, or right vowel signs and each segment is assigned with a label. The maximum number of segments in a given Malayalam character is not more than three segments and getting back the original character is based on the logic shown in the flow chart. SVM classifiers are used to label each of the segments V, 0, X and 1. The SVM (binary classifier) is applied to this multiclass character recognition problem by using one-versus-rest type method. The problem now is a n -class problem with n equal to the number of segments in total, but during training it was found that some of the confusion characters have similar appearance though this problem of similarity is well tackled by using the feature extraction approaches explained in the Section V. Hierarchical SVM classifiers are used with the reduction in the search space as shown in Figure 4.

VIII. RESULTS AND ANALYSIS

Malayalam was found most appropriate to evaluate the three methods of characterization to extract the extraordinarily distinct and dominant characteristic features. Feature vectors of all the 620 Malayalam characters including conjuncts extracted based on the methods outlined from the training data set. Around 1000 different samples of Malayalam words from different

textbooks and magazines were scanned to obtain the data set for training. The data for testing the approach was obtained by scanning over 100 different documents. The approach has been tested successfully and a good recognition rate is achieved in all the trials. The SVM classifier, is used to recognize the characters. A notable outcome of our approach is that it is inherently invariant to font face and proper normalization ensures invariance to font size. The paper has addressed the problem of character recognition through three structural approaches as features. Recognition of Malayalam characters based on distinct features has been demonstrated with good accuracy.

TABLE I.
RECOGNITION RATE OF DIFFERENT TYPES OF MALAYALAM CHARACTERS BASED ON THE PROPOSED METHODS

Malayalam characters	Absorption	Average Gap	Frequency Capture
General characters	93.47 %	94.16 %	95.31 %
Similar characters	84.23%	90.30 %	92.27 %
Long characters	86.21 %	92.21 %	94.54 %
Short characters	90.22 %	91.34 %	93.64 %
Conjunct characters	90.62 %	91.39 %	92.69 %

Confusion characters usually misclassify in classical approaches. It is thus appropriate to compare the proposed approaches on their performances in classifying such similar characters. It is fascinating to see that the frequency capture approach yields better results with good discrimination between similar characters.

The performance of the proposed methods in terms of the recognition rate are compared and is tabulated for five different categories of Malayalam characters on the basis of there column length and structure. It is clear from the Table I that the proposed methods perform well and appear promising for all types of printed Malayalam characters.

IX. CONCLUSION

The system was developed using C++ on Windows XP Platform. The proposed approaches have been tested successfully on the extended Malayalam character set. Several analyses have been performed to illustrate the viability of the proposed features using the three different structural approaches for classification. Identical characters were found to be far apart in feature space in all the three approaches. The methods were tested on approximately 6000 Malayalam characters and average recognition accuracy of 88.99% for Absorption, 91.88% for average gap and 93.69% for Frequency capture has been achieved.

The Figure 5 shows some examples of the words used for experimentation having the combination of left vowel

signs, consonants, right vowel signs, along with consonants which was successfully segmented and accurately classified using the approach explained in the Section V and Section VI for segmentation and feature extraction respectively.

മാല Maala
 തിര Tira
 ശില Shila
 കൂട Kuta
 ചുര Chura
 കൃതി Kruti
 ചെവി Chevi
 വേല Vela
 മൈന Maina
 തൊലി Toli
 കോടി Koti
 സൗമിനി Saumini

Figure5. Examples of Words used for Experimentation

ACKNOWLEDGEMENT

This work was supported in part by research grants from UGC for Major Research Project in Science and Technology, F.No. 32-113/2006

REFERENCES

- [1] Ashwin T V, P S Sastry "A font and size independent OCR system for printed Kannada documents using support vector machines", Saadhana, Vol. 27, Part 1, February 2002, pp. 35-58.
- [2] Pal U. , B. B. Chaudhuri 1997 Printed Devnagari Script OCR System. Vivek, vol.10, pp.12-24
- [3] Seethalakshmi R., Sreeranjani T.R., Balachandar T., Abnikant Singh, Markandey Singh, Ritwaj Ratan, Sarvesh Kumar 2005 "Optical Character Recognition for printed Tamil text using Unicode" Journal of Zhejiang University SCI 6A(11) pp.1297-1305
- [4] Negi Atul, Chakravarthy Bhagvati and.Krishna B 2001 An OCR system for Telugu. Proc. Of 6th Int. Conf. on Document Analysis and Recognition IEEE Comp. Soc. Press, USA,, pp. 1110-1114.
- [5] Lajish V. L., Suneesh T.K.K. and Narayanan N.K., "Recognition of Isolated Handwritten Character Images using Kolmogrov-Smirnov Statistical Classifier and k-nearest Neighbour classifier", Proc. Of the International Conference on Cognition and Recognition ICCR-05, Mandya, Karnataka, December, 2005
- [6] Renju John, G.Raju and D. S. Guru, "ID Wavelet Transform of Projection Profiles for Isolated Handwritten Malayalam Character Recognition", Proc. of International Conference on Computational Intelligence and Multimedia Applications 2007, Sivakashi, IEEE computer society Press, 2007, pp 481-485,.
- [7] P. S. Janardhanan. Issues in the development of OCR systems for Dravidian languages - proceedings of Akshara 94., BPB Publications, New Delhi, India 1994.
- [8] Malayalam standardization report May 2001
- [9] Burges C J C 1998 A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discovery 2: 955-974.