Preferred Computational Approaches for the Recognition of different Classes of Printed Malayalam Characters using Hierarchical SVM Classifiers

Bindu Philip
Department of Electronics & Communication
S J College of Engineering
Mysore - 570006

R. D. Sudhaker Samuel
Department of Electronics & Communication
S J College of Engineering
Mysore - 570006

ABSTRACT

Characterization of matrices for efficient classification has several options. There are various alternatives depending on the structure of the matrix. Different features can be adapted in different situations. Image recognition and in particular character recognition is an excellent example where large number of image matrices need to be stored and retrieved often at high speed, at the same time performing computational tasks, resulting in requirements of huge memory and computation time. Near 100% character segmentation accuracy is achieved based on a novel segmentation technique. Here feature extraction is based on the distinctive structural features of machine-printed text lines in these scripts. The final recognition is achieved through Support Vector Machine (SVM) classifiers. The proposed algorithms have been tested on a variety of printed Malayalam documents. Recognition rates between 97.72% and 98.78% have resulted.

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

Pattern recognition, character classification, segmentation, Optical character recognition, Singular value decomposition, marginal frequency, Support vector machine classifier, Malayalam OCR

1. INTRODUCTION

In the recent past OCRs for Indian languages has become an active area of research owing to their application in the digitization of vast amount of literature present in print media. Research in the field of recognition of Indian script posses several problems mainly due to the large set of characters in the alphabet, their complex shapes and sizes. OCR technology converts images of machine-printed characters into machine-readable characters. OCR systems aim at enabling machines to recognize optical symbols without human intervention [8]. Character extraction and recognition techniques have potential application in any domain where massive document image-bearing texts must be interpreted, analyzed and processed [12]. Selection of the segmentation strategy, features, and classifier are very important steps in the construction of an OCR. It is a well established fact that a rich feature space can solve the character classification problems with reasonable success rate. Malayalam script is rich in structural features and indicating the probable successes of structural approaches for feature extraction for the efficient representation of Malayalam characters described in section 6. In Malayalam script the space between the sub-characters and the core character is equal and same as the space between the characters within a word rendering the character segmentation process quite complex as conventional profiling methods fail. This paper presents a novel segmentation technique with near 100% character segmentation accuracy and is explained in section 5. SVM classifier is designed for classification in order to classify large data set of 1072 Malayalam characters and is discussed in section 7.

2. SOME RECENT DEVELOPMENTS IN OCR TECHNIQUES FOR INDIAN SCRIPTS

Some of the existing techniques used in OCR for Indian scripts is presented here. Recognition and a review of the OCR work done on Indian language are excellently reviewed by Pal & Chaudhuri [8]. An OCR for Telugu is reported by Negi, *et. al*. [9] here instead of segmenting the words into characters as usually done, words are split into connected components (glyphs). An attempt to recognize Telugu script using KNN and Fringe distance is reported in [9]. 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 [5]; Seethalakshmi, *et. al*, [10], reported a Tamil OCR using Unicode and SVM classifier. Shivsubramani K *et al*. [11] gives an efficient method for recognizing printed Tamil characters exploring the interclass relationship between them and they accomplished using Multiclass Hierarchical Support

Vector Machines, a new variant of Multi Class Support Vector Machine which constructs a hyperplane that separates each class of data from other classes. Renju John, *et.al.,* [13] 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*., [14]. A comprehensive study on the success rate of well known feature extraction methods in terms of recognition accuracy and computational complexity for printed Malayalam characters is yet to be reported.

3. 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 left, 2 right vowel signs, 30 commonly used conjuncts, 36 consonants and the vowel signs are as shown in Figure 1. The dependent vowels do are depicted in combination with a consonant or consonant cluster [15]. Explicit appearance of a dependent vowel in a syllable overrides the inherent vowel of a single consonant character. The independent vowels are used to write syllables which start with a vowel. 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 [16] as shown in Figure 1.

Vowel Sign	Left/Right of the Consonant/ conjunct	Vowel sign attached to ക (ka) (example)	
ാ	Right	കാ	kA
ി	Right	കി	Ki
ീ	Right	കീ	kI
ു	Right	കു	Ku
ൂ	Right	കൂ	kU
ൃ	Right	കൃ	kRu
െ	Left	കെ	ke
േ	Left	കേ	kE
ൈ	Left	കൈ	kai
ൊ	Left & Right	കൊ	ko
ോ	Left & Right	കോ	kO
ൗ	Right	കൗ	kau

Figure 1: Malayalam Vowel Signs

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 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.

4. 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 profile technique. The characters and sub characters in printed Malayalam text have uniform distance of separation within a word and thus segmentation of full characters in a Malayalam word is a great challenge. Character segmentation is thus done using a novel segmentation algorithm as explained in Section 5. The segmented fragments are now normalized to a height of ***m1*** pixels preserving the length of the characters. It is significant to mention that ***m1 = 50*** was found to be an optimum value after several trials. The problem now reduces to the characterization of ***m1xn*** image matrices. The value of ***m1*** 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. In the classification stage Hierarchal SVM classifiers are used with the reduction in the search space which is explained in Section 7.

5. 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.

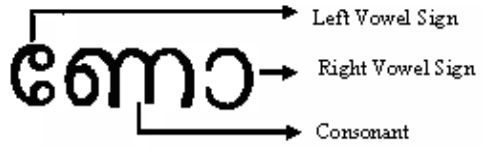

Figure 2: A Typical Malayalam Character

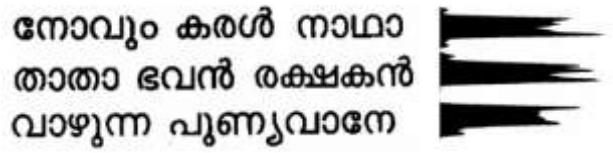

Figure 3: Example of scanned Malayalam printed Text and its Horizontal Projection Profiles

Line segmentation is performed using classical horizontal projection profile technique as shown in Figure 3 and the end of the word is identified by a larger valley in the vertical projection profile as shown in Figure 4. Note that the sub-characters are extracted using the smaller valley of projections.

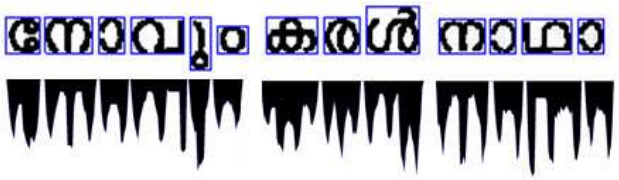

Figure 4: Vertical Projection Profiles and Segmentation result at sub character level

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), VSL represents the vowel signs to the left (2 in all) while VSR represents the vowel signs to the right (7 in all). 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.

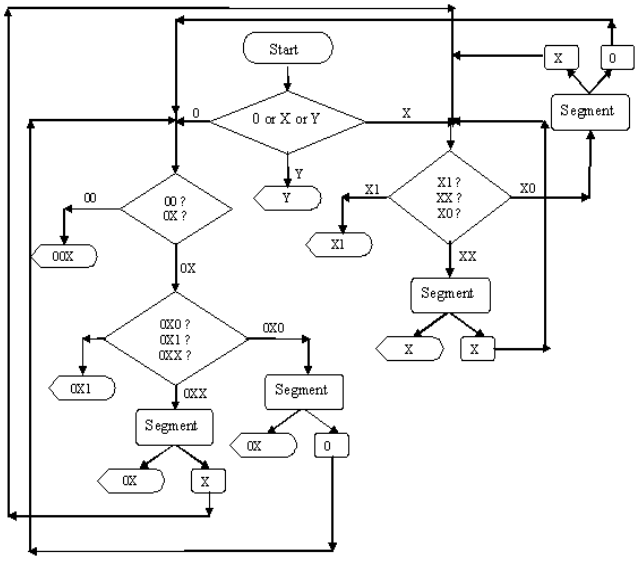

Figure 5: Segmentation Logic Flow Chart

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.

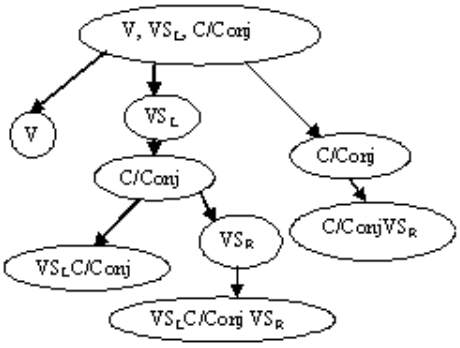

Figure 6: Classification Search Space

6. FEATURE EXTRACTION

Feature extraction is the identification of appropriate measures to characterize the component images distinctly. There are many popular methods to extract features. Considering the entire image as the feature is at one end of the spectrum. This representation is bulky and contains redundant information. In [8] a survey on feature extraction methods for character recognition is reviewed. Feature extraction method includes Geometric moment invariants, Zernike Moments, Template matching, Deformable templates, Projection Histograms, Contour profiles, Zoning, Spline curve approximation, Unitary Image transforms, Graph description, Fourier descriptors , Gradient feature and Gabor feature. This paper presents the characterization of ***mxn*** matrices using three rather diverse methods, the average gap analysis, the singular values and marginal frequency capture to form distinct features tested on the classification of the complete extended character set of the Malayalam language.

6.1 Average Gap Analysis

The average gap along each row is computed to get feature vector $x \in \mathbb{R}^m$ of the matrix $A \in \mathbb{R}^{mxn}$ given by equation

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

$\sigma_i = (\sum_{j=1}^{n+1} | b_{i,j+1} - b_{i,j} |)/2$ with each row padded with 1s on either ends for the sake for computational convenience, where $(b_{ij}) = (a_{ij})$, $(b_{i,1}) = (b_{i,n+2}) = 1 \forall i$ for i =1…m and j = 2…n+1

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

6.2 Marginal Frequency capture

This method essentially captures the frequency of transitions along each row, to form the feature vector, $x \in \mathbb{R}^m$ of the matrix $A \in \mathbb{R}^{mxn}$, given by equation (2)

$$x_i = \frac{\sum_j a_{ij}}{\sum_j \sum_i a_{ij}} \qquad (2)$$

Corresponding to α images, there would be α image feature vectors given by $x^k \in \mathbb{R}^m$ where, k=1,2,….. α. It is rather important to mention that there are several problem spaces where rich information along rows useful in classifying images with small differences would be lost in the process of squaring a rectangular matrix in the course of normalization.

6.3 Singular Values

In this approach too, the row information is captured during the process of arriving at the singular values [1].

Now, if $A \in \mathbb{R}^{mxn}$, then $\exists$ Orthogonal matrix $U \in \mathbb{R}^{mxm}$ and $V \in \mathbb{R}^{nxn}$ ∋

$U^T AV = diag(\sigma_1, ..., \sigma_s) \in \mathbb{R}^{mxn}$

where $s = \min\{m, n\}$ and $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_s \geq 0$.

For the extended Malayalam character set considered here, $s = m$ always for all images. These singular values are used to define a feature vector $\hat{x} \in \mathbb{R}^m$ defined by $\hat{x}_i = \sigma_i$, where $i = 1, 2, ...m$.

Let $\sigma_i = \{\sigma_1, \sigma_2, ...\sigma_p, \sigma_{p+1}, ...\sigma_m\}$.

In our examples it is seen that $\sigma_i \gg \sigma_j$ for $i = 1, 2, ...p$ and $j = p+1, ...m$, implying that there are few (p) dominant singular values. It is found experimentally that these p dominant singular values are sufficient to characterize the image matrix. The feature vector defined here now becomes

$x_i = \sigma_i$, where $i = 1, 2, ...p$.

This definition of dominant singular values is further justified by the fact that $\left| \|x\| - \|\hat{x}\| \right| \leq \xi$, $\hat{x} \in \mathbb{R}^m$ and $x \in \mathbb{R}^p$, ξ being small, implying that there is insignificant contribution to the feature of the image from the ***(m-p)*** singular values not being considered. The value of ***p*** is chosen by defining a value of ξ for reliable and robust classification. Five dominant singular values are selected as features for distinct classification of the characters. Any further increase in the number of singular values did not improve the recognition rate which was 95.51% while considering the general characters.

7. SUPPORT VECTOR MACHINE (SVM) CLASSIFIER

The Support Vector Machine 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). 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 principle of an SVM is to map the input data onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyperplane with maximum margin between the two classes in the feature space[6].

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

$u_k = 1$ if $x_k \in \omega_1$

$u_k = -1$ 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.

$x_k \rightarrow \Phi(x_k)$ and $x_j \rightarrow \Phi(x_j)$

$\Rightarrow x_j^T x_k \rightarrow \Phi(x_k)^T \Phi(x_j)$ the function Φ is replaced by a symmetric and separable function called kernel Δ.

The Kernel Δ Function is defined as $\Delta\ x_k, x_j = \exp\ -\alpha \| x_k - x_j \|^2$

$\Delta : \square^n X \square^n \rightarrow \square, \Delta\ x_k, x_j = \Phi\ x_k\ ^T \Phi\ x_j$

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

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

7.1 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 6. Hierarchal SVM classifiers are used with the reduction in the search space as shown in Figure 4.

8. RESULTS AND ANALYSIS

Malayalam was found most appropriate to evaluate these methods of characterization to extract the extraordinarily distinct and dominant characteristic features. A database is created with complete set of Malayalam characters including complex extended characters and conjugate characters. A training set is formed for all the characters in the database and the feature of each is extracted based on the methods outlined. Feature vectors of all the 1072 extended Malayalam characters form the training set in each database. Three types of features were extracted based on the methods proposed. The average gap and marginal frequency methods are referred to as structural methods as they capture the row information. The feature vector is of dimension 50.

Table 1: Analysis of features of short characters

Sl No	Characters	Average Gap	Marginal Frequency	SVD
1.	(ma)	0.159	0.018	0.202
2.	(da)	0.111	0.012	0.280
3.	(bha)	0.132	0.063	0.188
4.	(rxa)	0.122	0.072	0.129
5.	(zha)	0.124	0.043	0.262

The SVD technique considered here gave further dimensionality reduction with a feature vector of dimension 5 as further increase in the features did not give any improvement in the recognition rate. The extraction of dominant singular values is however computationally more expensive than the other two proposed methods. Table 1 shows the distance between characters under test and the characters used in training, averaging over 200 different samples. Table 2 shows the similar data for long characters. Table 3 however shows the distance between similar characters by all the three methods.

Table 2: Analysis of features of long characters

Sl no	Characters	Average Gap	Marginal Frequency	SVD
1.	(jho)	0.073	0.102	0.280
2.	(ghai)	0.053	0.263	0.062
3.	(Dhai)	0.044	0.165	0.049
4.	(jhO)	0.021	0.086	0.127
5.	(sai)	0.075	0.128	0.268

The data in Table 1, 2 and 3 indicate that feature extraction based on average gap along rows is the best suited for long characters while features based on marginal frequencies give good results for short characters. The dominant singular values perform well in-between as well as in the case of confusing characters.

Table 3: Analysis of features of confusion characters

Sl No	Characters		Average Gap	Marginal Frequency	SVD
1.	(da)	(Ba)	0.221	0.307	3.63
2.	(va)	(Ka)	0.231	0.314	2.34
3.	(chu)	(choo)	0.256	0.332	2.37
4.	(da)	(edha)	0.488	0.461	5.13
5.	(e)	(ee)	0.299	0.159	3.72

Table 4: Recognition rate of the proposed methods in for the different types of Malayalam characters experimented

Methods / Characters	Average Gap	Marginal Frequency	SVD
General characters	97.15%	97.72%	95.51%
Similar characters	94.2%	95.4%	97.34%
Long characters	98.61%	96.31%	95.11%
Short characters	95.12%	98.56%	95.61%
Conjunct characters	96.75%	95.88%	98.78%

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 proposed approaches yields better results with good discrimination between similar characters. The performances are compared and their accuracy is tabulated for five different types of Malayalam characters on the basis of their length and structure. It is clear from the Table 3 that the proposed methods perform well and appear promising for different classes of printed Malayalam characters. SVM classifier is used for recognizing the segments in each full character. SVMlight [17] package is used to train the SVM classifier. A hierarchical SVM for classifying the characters as discussed in Section 7.1.

9. 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 suitability of the proposed features for different character lengths and situations as illustrated through the tabulations. The performances may be classified as shown in Table 5. Column lengths are based on a scanning resolution of 300 dpi.

Table 5: Character type and its best suitable feature vector

Sl. No	Character Type	Description	Best Feature
1.	Short characters	Column lengths less than 80	Marginal frequency
2.	Long characters	Column lengths greater than 200	Average gap
3.	Mid characters	Column lengths ranging between 80 and 200	Marginal frequencies and average gap
4.	Confusing characters	Similar looking characters with very minor variations	Dominant singular values
5.	Conjunct characters	Complex characters made up of two or more consonants or vowels	Dominant singular values

10. 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.

11. REFERENCES

[1] Golub G.H. and Loan V.C.F. Matrix Computations. The John Hopkins Press, 1989.

[2] Eldén L Numerical linear algebra in data mining. Acta Numerica 15, 327-384, Cambridge University Press, 5/2006

[3] Intel Open Source Computer Vision Library http://www. intel.com/technology/computing/opencv/index.htm

[4] Aparna K G, Ramakrishnan A G, "A complete Tamil Optical Character Recognition System",5th International Workshop on Document Analysis Systems DAS 2002, Princeton, NJ, USA,2002, pp. 53-57.

[5] Ashwin T V, P S Sastry "A font and size independent OCR system for printed Kannada documents using support vector machines", Sadhana, Vol. 27, Part 1, February 2002, pp. 35–58.

[6] Burges C. J. C, .A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, pp 121-167.

[7] Jain A.K and Taxt T, "Feature extraction methods for character recognition-A Survey", Pattern Recognition, vol. 29, no. 4, pp. 641-662., 1996,

[8] Pal.U, Chaudhuri B. B, "Indian Script Character recognition: A survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[9] 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.

[10] Seethalakshmi R., Sreeranjani T.R., Balachandar T., Singh A, Singh M, Ratan R, Kumar S, 2005 "Optical Character Recognition for printed Tamil text using Unicode" Journal of Zhejiang University SCI 6A(11) pp.1297-1305

[11] Shivsubramani K, Loganathan R, Srinivasan CJ, Ajay V, Soman KP "Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters" twentieth international conference on artificial intelligence, Hyderabad, India, January 2007, pp 93-97.

[12] Jawahar C. V., Pavan K M. N. S. S. K, and S. S. Kiran R, "Recognition of Indian Language Characters using Support Vectors Machines," Technical Report TR-CVIT-22, International Institute of Information Technology, Hyderabad, 2002.

[13] John R, Raju G and Guru D. S, "1D 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,.

[14] 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

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

[16] Malayalam standardization report May 2001.

[17] Joachims T 1999 SVMlight. http://www-ai.informatik.uni-dortmund.de/forschung/verfahren/svmlight/svmlight.eng.html