I.J. Image, Graphics and Signal Processing, **2013, 4, 53-59**
Published Online April 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2013.04.07

A System for Offline Recognition of Handwritten Characters in Malayalam Script

Jomy John, Kannan Balakrishnan, Pramod K. V
Department of Computer Applications
Cochin University of Science and Technology, Kochi, Kerala, India

Abstract— In this paper, we propose a handwritten character recognition system for Malayalam language. The feature extraction phase consists of gradient and curvature calculation and dimensionality reduction using Principal Component Analysis. Directional information from the arc tangent of gradient is used as gradient feature. Strength of gradient in curvature direction is used as the curvature feature. The proposed system uses a combination of gradient and curvature feature in reduced dimension as the feature vector. For classification, discriminative power of Support Vector Machine (SVM) is evaluated. The results reveal that SVM with Radial Basis Function (RBF) kernel yield the best performance with 96.28% and 97.96% of accuracy in two different datasets. This is the highest accuracy ever reported on these datasets.

Index Terms— Character recognition, Malayalam, Gradient, Curvature, Principal Component Analysis, SVM, RBF

I. INTRODUCTION

Character recognition has been one of the most active topics in pattern recognition for several decades. It is the process of converting an image representation of a document into digital form. The document image can be printed or handwritten. Handwritten data is converted to digital form either by scanning the writings on paper or by writing with a special pen on an electronic surface such as a digitizer combined with a liquid crystal display. The two approaches are termed as off-line and on-line handwriting, respectively. The order of strokes made by the writer is available in the latter but only the completed image is available in the former [1]. In this paper, we propose an offline recognition system for Malayalam character images.

Malayalam character recognition is a difficult problem due to a very large number of characters and many instances of highly similar characters. The problem becomes more difficult in the handwritten domain due to the variability writing style of each individual. Malayalam is one of twenty two scheduled languages of India, with rich literary heritage [2]. It has official language status in the State of Kerala and Union territories of Lakshadweep and Mahe. The language is spoken by around 35 million people and it is ranked eighth in terms of the number of speakers in India. Malayalam script is derived from the Grantha script, an inheritor of olden Brahmi script [3]. It is in close proximity to Tamil and has indelible impression of Sanskrit. It also has the influence of Arabic. Consequently, Malayalam language is enriched with largest number of characters among all Indian languages. Most of Malayalam alphabets have circular shapes. It is phonetic because words are written exactly as they are pronounced. It is syllabic in nature and alphabets are classified into vowels and consonants. The script consists of 15 vowels and 36 consonants, called basic characters. In addition to this basic set, the script contains vowel modifiers, half consonants and a large number of compound characters. The script also contains 10 numerals, but it is seldom in use. Instead Arabic numerals are used in practice. The set of vowels and consonants are depicted in Fig. 1.

A handwritten character recognition system consists of four phases: pre-processing, feature extraction, classification and post-processing. Among them, feature extraction is one of the most important factors in achieving high recognition performance. In this work, the feature extraction phase consists of gradient and curvature calculation and dimensionality reduction using Principal Component Analysis. Directional information from the arc tangent of gradient is used as gradient feature. Strength of gradient in the quantised curvature direction is used as the curvature feature. The proposed system uses a combination of gradient and curvature feature in reduced dimension as the feature vector. For classification, discriminative power of SVM is used. The organization of this paper is as follows: the next section discusses feature extraction using gradient and curvature of character image; section 3 introduces SVM, section 4 evaluates the system with experimental results and section 5 concludes the paper.

II. FEATURE EXTRACTION

A. Gradient Computation

The gradient of an image f is defined as a vector $[g_x\ g_y]^T$ which points to the direction of the greatest rate of change of f at location (x, y). It can be approximated by Roberts diagonal gradient operator as in (1)

vowels				consonants									
അ	ആ	ഇ	ഈ	ക	ഖ	ഗ	ഘ	ങ	ച	ഛ	ജ	ഝ	ഞ
ഉ	ഊ	ഋ	എ	ട	ഠ	ഡ	ഢ	ണ	ത	ഥ	ദ	ധ	ന
ഏ	ഐ	ഒ	ഓ	പ	ഫ	ബ	ഭ	മ	യ	ര	ല	വ	ശ
ഔ	അം	അഃ		ഷ	സ	ഹ	ള	ഴ	റ				

Fig. 1 Vowels and consonants of Malayalam language

$$g_x = f(x+1, y+1) - f(x, y)$$
$$g_y = f(x+1, y) - f(x, y+1) \quad (1)$$

Strength (G) and direction (φ) of gradient at location (x, y) is computed as in (2),

$$\varphi = \tan^{-1}\left[\frac{g_y}{g_x}\right]$$

$$G = \sqrt{g_x + g_y} \quad (2)$$

B. Curvature Computation

Curvature feature used in this paper is calculated using bi-quadratic interpolation method [4]. The curvature c at x_0 is defined as

$$c = \frac{y''}{\sqrt{(1+y'^2)^3}} \quad (3)$$

where y=g(x) is the equi – gray scale curve passing through x_0, (x, y) is the spatial coordinates of x, y' and y'' are the first and the second order derivatives of y respectively. The derivatives of y' and y'' are derived from bi-quadratic interpolating surface for the gray scale values in the 8-neighborhood of x_0. Fig. 2(a) shows the 8-neighbors of x_0 and 2(b) shows the pixel values in the neighborhood where f_k denote the pixel value at x_k

x_4	x_3	x_2
x_5	x_0	x_1
x_6	x_7	x_8

f_4	f_3	f_2
f_5	f_0	f_1
f_6	f_7	f_8

Fig. 2(a) 8-neighbours of x_0 (b) pixel values in the neighborhood

The bi-quadratic surface is given by

$$z = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 1 \\ y \\ y^2 \end{bmatrix} \quad (4)$$

The curve passing through x_0 is defined as,

$$f_0 = a_{00} + a_{10}x + a_{20}x^2 + y(a_{01} + a_{11}x + a_{21}x^2) + y^2(a_{02} + a_{12}x + a_{22}x^2) \quad (5)$$

where f_0 is the pixel value at x_0

Differentiating both sides with respect to x and substituting the value (0, 0) of x_0, the values of y' and y'' at x_0 are given by

$$y' = -a_{10} / a_{01}$$
$$y'' = -2(a_{10}^2 a_{02} - a_{01}a_{10}a_{11} + a_{01}^2 a_{20}) / a_{01}^3 \quad (6)$$

Solving for 8-neighbours of x_0, the coefficients of the bi-quadratic surface are given by,

$$a_{10} = (f_1 - f_5)/2,$$
$$a_{20} = (f_1 + f_5 - 2f_0)/2,$$
$$a_{01} = (f_3 - f_7)/2,$$
$$a_{02} = (f_3 + f_7 - 2f_0)/2,$$
$$a_{11} = (f_2 - f_8) - (f_4 - f_6)/4, \quad (7)$$

Substituting the values of y' and y'' in (3), the curvature can be calculated as

$$c = -2(a_{10}^2 a_{02} - a_{01}a_{10}a_{11} + a_{01}^2 a_{20}) / (a_{10}^2 + a_{01}^2)^{3/2} \quad (8)$$

C. Principal Component Analysis

PCA is one of the simplest and most robust ways to reduce the dimensionality of a data set. This reduction is achieved by transforming the data set to a new set of variables, called principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in the original set of variables [5]. Hence after PCA, the feature set can be truncated without losing too much information.

D. Feature set creation

Step 1: Resize character image to 100 x 100

Step 2: Normalize the input image so that mean value is zero and maximum value is one

Step 3: Divide the input image into 10 horizontal and 10 vertical blocks.

Step 4: Compute gradient image and curvature image as in Section 3.1 and 3.2 Quantize the direction of gradient and curvature image into 32 equi-length ranges

Step 5: Calculate

(i) the strength of gradient accumulated in the 32 gradient direction in each block using (2).
(ii) the strength of gradient accumulated in the 32 curvature direction in each block using (8).

Step 6: Reduce the spatial resolution from 10x10 to 5x5 by down sampling every two horizontal and every two vertical blocks and reduce the directional resolution from 32 to 16 each with Gaussian filter for the image in 5(i). This produces a feature vector of size 400 (5 vertical, 5 horizontal and 16 directional resolutions). Apply a variable transformation $y = x^{0.4}$ to make distribution as Gaussian

Step 7: Repeat Step 6 for the image in 5(ii)

Step 8: Construct feature set

FS # 1 as in 6 of dimension 400
FS # 2 as in 7 of dimension 400
FS # 3 by concatenating 6 and 7 of dimension 800

Step 9: Repeat step 1 to step 8 for each image in the dataset.

Reduce the dimensionality of each feature set (FS#1, FS#2 and FS # 3) without compromising accuracy, using principal component analysis. Fig. 3 displays the handwritten character image ‘ah’ (ആ) at different stages of feature extraction.

III. CLASSIFICATION

Support vector machines have proven to have good performance in handwritten character recognition problems and is considered to be the state-of-the-art tool for linear and non-linear classification [6]. Liu et al.[7] have evaluated the performance of several classifiers including SVM on CENPARMI, CEDAR and MNIST handwritten numeral databases. Bellili et al. [8] introduced a hybrid Multilayer Perceptron (MLP) and SVM classifiers for handwritten digit recognition. Handwritten Tamil Character Recognition system using SVM classifiers were proposed by Shanthi et al. [9]. Bhowmik et al. [10] provides SVM based hierarchical classification schemes for recognition of handwritten Bangla characters. To the best of our knowledge, the use of SVM in offline handwritten Malayalam characters was reported only in our previous paper [11].

SVM belongs to the class of supervised learning algorithms, based on statistical learning theory. This classifier has been originally proposed for binary classification in literature and learning algorithm comes from an optimal separating hyper-plane, developed by Vapnik [6]. For binary classification, a linear decision function $f(x) = w^T x + b$ is used, where w is the weight vector and b is a bias. Classification is given by sign of $f(x)$, which can be -1 or +1. The optimal solution is obtained when this hyper plane is located in the middle of the two classes. The points that constrain the width of the margin are called support vectors. Support vectors require the solution to the optimization problem of (9), where X_i are instances and y_i are class labels.

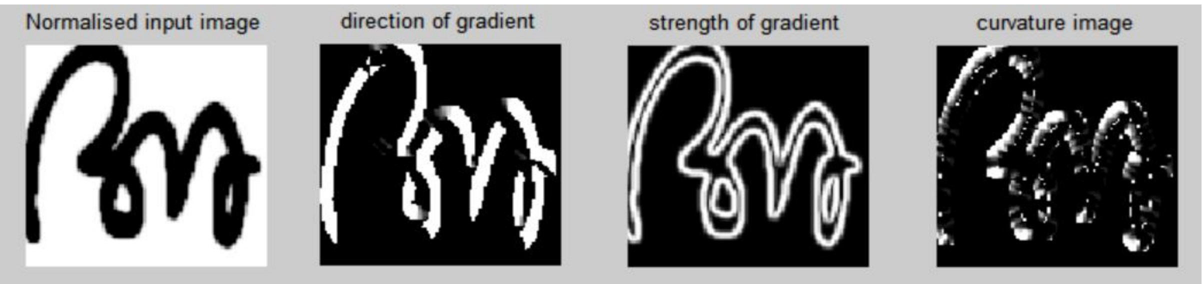

Fig. 3 Input image ‘ah’ (ആ) at different stages of feature extraction

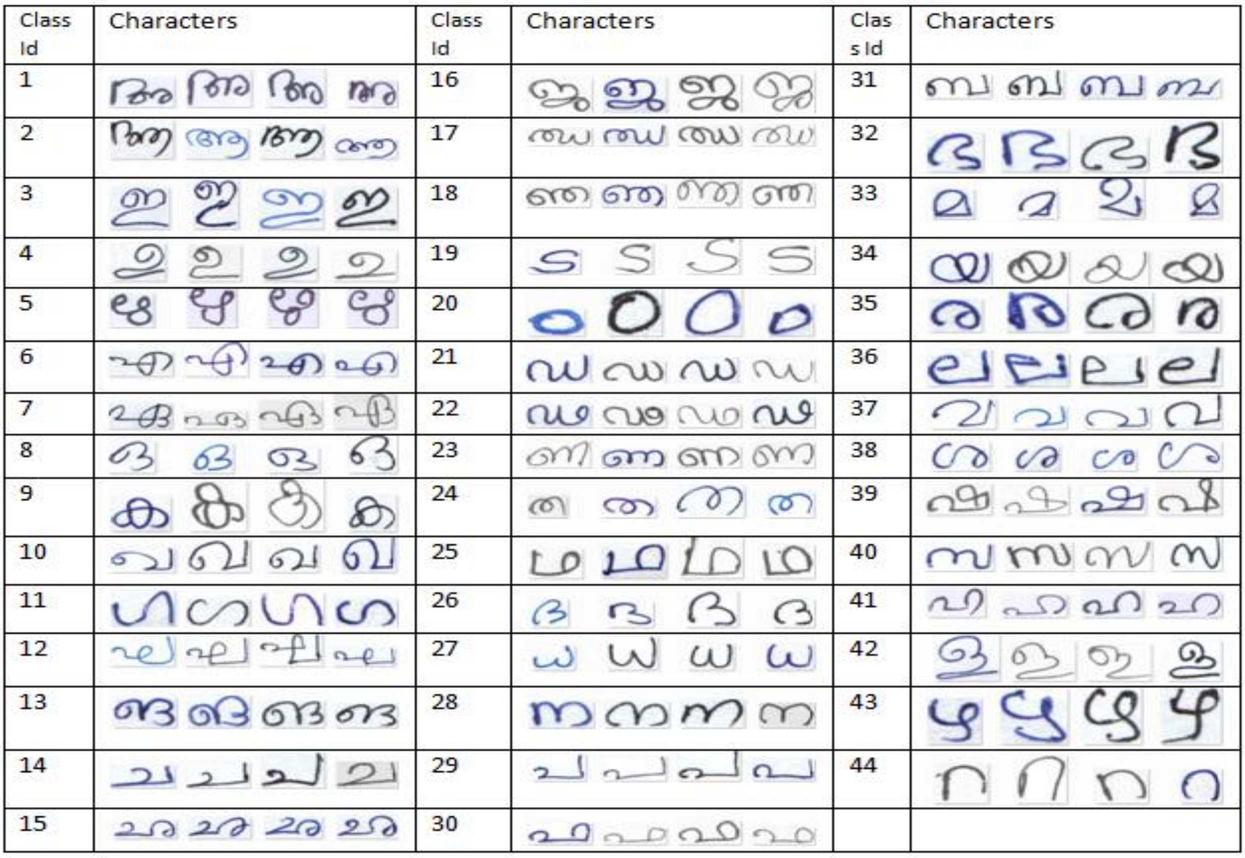

Fig. 4 Samples of 44 basic Malayalam characters from data set I

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

subject to $(y_i w^T \phi(X_i) + b) \geq 1 - \xi_i$, $\xi_i \geq 0,$ (9)

where C is the soft margin parameter and ξ_i is a slack variable. In the case of linearly inseparable feature space, the training vectors X_i are mapped into a higher dimensional space by the function ϕ. $K(X_i, X_j) \equiv \phi(X_i)^T \phi(X_j)$ is termed as the kernel function. We have used RBF kernel in which $K(X_i, X_j) = \exp(-\gamma \| X_i - X_j \|^2), \gamma > 0$ where γ is a kernel parameter.

IV. Experimental results and discussion

Experiments are done on two datasets namely, data set I and data set II. Stratified 10 fold cross-validation is used for all the experiment. Here database is divided into 10 subsets and testing is done on each subset using other nine subsets in learning. The recognition rates for all the test subsets are averaged to calculate recognition accuracy. For classification, support vector classifier with RBF kernel is selected. The parameter γ and C are set to 0.02 and 100 respectively. The parameters values are chosen based on several experiments.

Data Set I: Data was collected from different persons of the population in Kerala, including different age groups without imposing any constraints. It represents wide variety of writing styles. Digitization of collected samples are done by a Flat-bed scanner (manufactured by HP, Model Name: Scanjet 2400), by setting dpi to 300. The characters are segmented using connected component analysis. The database contains 10,946 handwritten isolated Malayalam characters of 44 basic (vowels as well as consonants) alphabets. Fig. 4 displays samples of all these characters. The character images are converted to grayscale before feature extraction.

The classifier gives a recognition result of 95.65% using FS#1 and 92.06% using FS#2. The dimensionality of each is reduced from 400 to 200. But using FS#3 as a feature vector, the classifier yields a result of 96.28%. Here the dimension is reduced from 800 to 200. These results show significant improvement to all our previous works [11-13]. Fig. 5 displays a plot of confusion matrix.

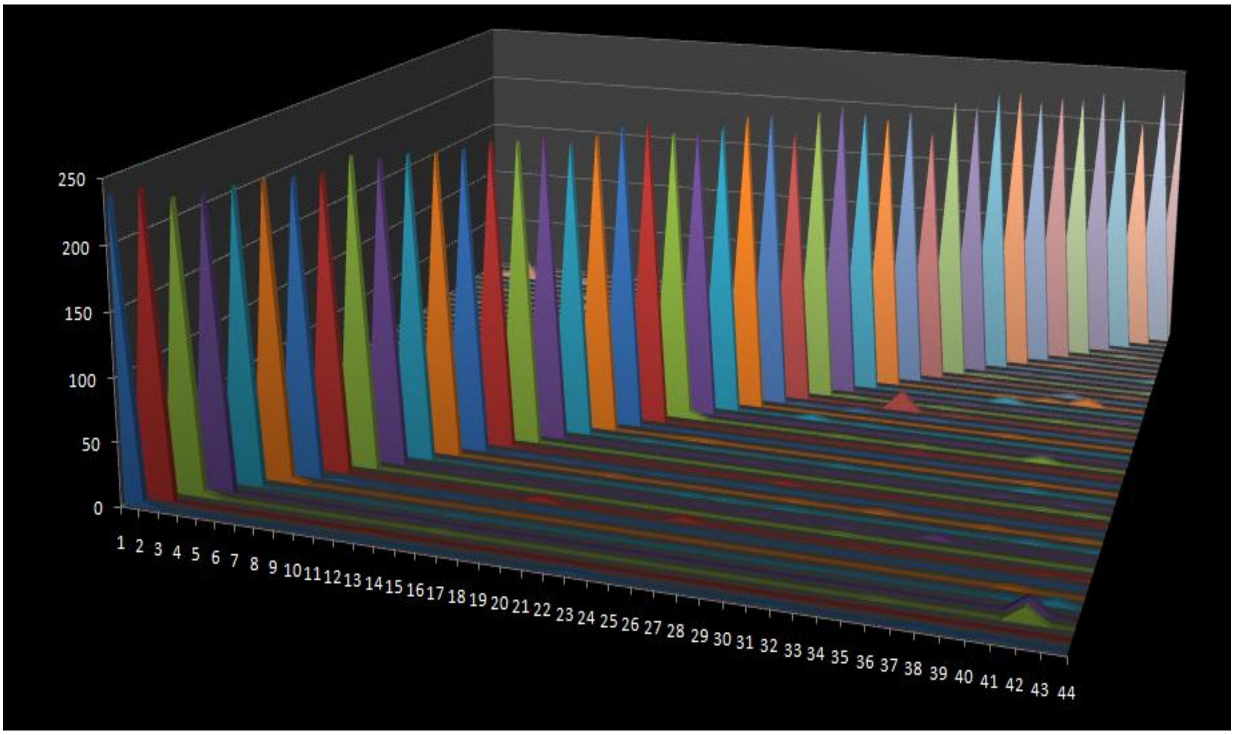

Fig. 5 Plot of confusion matrix with FS#3 on data set I

Data set II: As the recognition accuracy a system depends also on a set of factors including, the number of classes, number of samples used per class, way in which samples are collected etc., we have repeated the experiment on the dataset as specified in [14, 15], in order to compare the recognition result of others works. This dataset contain 13200 isolated handwritten Malayalam characters in binary form. A 2x2 averaging filter is applied to convert these images to gray scale.

We have chosen FS#3 as the feature set due to its better performance compared to FS#1 and FS#2. The dimension of this feature vector can be reduced from 800 to 400 without compromising the accuracy. The classifier produces a recognition result of 97.96% in this data set. Correctly classified instances for each class are depicted in Table 1. In recent works on similar problem, BP Chacko obtained an accuracy of 83.98% and 96.83% with division point features using a hybrid learning algorithm and online sequential extreme learning machine as in [14, 15] respectively. In[16] , BS Moni used direction codes as features and obtained an accuracy of 95.4%. The recognition result of 97.96% by our method is highest rate ever reported on this database. Even if we reduce the dimension to just 45, the accuracy is diminished by only 0.7045%. Table 2 displays the recognition accuracy for different dimension of features for this feature set.

Table 1. Individual classification results with FS#3 on data set 2

Class	Character	Correctly classified	Class	Character	Correctly classified	Class	Character	Correctly classified	Class	Character	Correctly classified
1	അ	294	12	ഘ	298	23	ണ	298	34	യ	297
2	ആ	293	13	ങ	293	24	ത	294	35	ര	299
3	ഇ	282	14	ച	298	25	ഥ	297	36	ല	294
4	ഉ	285	15	ഛ	299	26	ദ	290	37	വ	294
5	ഋ	292	16	ജ	296	27	ധ	293	38	ശ	294
6	എ	295	17	ഝ	293	28	ന	295	39	ഷ	289
7	ഏ	298	18	ഞ	289	29	പ	297	40	സ	298
8	ഒ	295	19	ട	298	30	ഫ	291	41	ഹ	294
9	ക	295	20	ഠ	298	31	ബ	296	42	ള	281
10	ഖ	294	21	ഡ	296	32	ഭ	288	43	ഴ	292
11	ഗ	291	22	ഢ	296	33	മ	295	44	റ	297

Table 2. Accuracy vs. feature dimension with FS#3 on data set 2

Accuracy (%)	97.9621	97.9242	97.7500	97.3636	97.2576
Feature dimension	400	200	100	50	45

Moreover, if we analyse the result on data set I as well as on data set II, we can understand that the error is just due to the high similarity of handwritten characters. Class 1 (അ) with class 2 (ആ); class 3 (ഇ) with class 4 (ഉ) and 42 (ള); class 5 (ഊ) with class 43 (ഴ) and so on. Even though these characters are clearly distinguishable in printed format, it creates confusion in handwritten domain. If we ignore the errors between very similar characters, the recognition rate can be enhanced to a great extent.

V. Conclusion

In this paper, we propose a character recognition methodology for Malayalam handwritten character images. We have used efficient feature extraction method using gradient and curvature features. As the characters are curved in nature, combination of gradient and curvature feature has proven to yield the best recognition accuracy on these datasets. The only problem is in discriminating similar characters. So our future research is focused on identifying more robust features powerful enough to separate similar character pairs.

Acknowledgements

The authors are grateful to Dr. Babu Anto P., Professor, Department of Information Technology, University of Kannur, for providing data set II, for research purpose. The first author is thankful to University Grants Commission for providing fellowship to carry out research work.

References

[1] Plamondon, R. and S.N. Srihari, *Online and off-line handwriting recognition: a comprehensive survey*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000. **22**(1): p. 63-84.

[2] Govindaraju, V. and S. Setlur, *Guide to OCR for Indic Scripts: Document Recognition and Retrieval*. 2009: Springer Publishing Company, Incorporated. 325.

[3] Ghosh, D., T. Dube, and A.P. Shivaprasad, *Script Recognition;A Review*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010. **32**(12): p. 2142-2161.

[4] Shi, M., et al., *Handwritten numeral recognition using gradient and curvature of gray scale image*. Pattern Recognition, 2002. **35**(10): p. 2051-2059.

[5] Jolliffe, I.T., *Principal Component Analysis*. 2002: Springer.

[6] N, V.V., *The Nature of Statistical Learning Theory*, 1999, Springer, Information science and statistics: Berlin.

[7] Liu C L, N.K., Sako H, Fujisawa H, *Handwritten digit recognition: Benchmarking of state-of-the-art techniques*. Pattern Recognition. **36**(10): p. 2271-2285.

[8] Bellili A, G.M., Gallinari P, *An MLP-SVM combination architecture for offline handwritten digit recogniton:reduction of recogniton errors by support vector machine rejection mechanisms*. International Journal of Document Analysis and Recogniton, 2003. **5**: p. 244-252.

[9] Shanthi, N. and K. Duraiswamy, *A novel SVM-based handwritten Tamil character recognition system*. Pattern Analysis and Applications, 2010. **13**(2): p. 173-180.

[10] Bhowmik T K, G.P., A. Roy, Parui S. K, *SVM-based hierarchical architectures for handwritten Bangla character recognition*. International Journal of Document Analysis and Recognition, 2009.

[11] John, J., K.V. Pramod, and K. Balakrishnan, *Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier*. Procedia Engineering, 2012. 30(0): p. 598-605.

[12] John, J., K.V. Pramod, and K. Balakrishnan. *Offline handwritten Malayalam Character Recognition based on chain code histogram*. in *Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on*. 2011.

[13] Joseph, S., et al. *Content based image retrieval system for Malayalam handwritten characters*. in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*. 2011.

[14] Chacko, B.P. and P.B. Anto, *A Hybrid Learning Algorithm for Handwriting Recognition in Information Systems for Indian Languages*, C. Singh, et al., Editors. 2011, Springer Berlin Heidelberg. p. 232-235.

[15] Chacko, B.P. and A.P. Babu. *Online sequential extreme learning machine based handwritten character recognition*. in *Students' Technology Symposium (TechSym), 2011 IEEE*. 2011.

[16] Moni, B.S. and G. Raju, *Modified quadratic classifier and directional features for handwritten Malayalam character recognition*. Int. J. Comput. Appl, 2011: p. 30-34.

Jomy John, received her M. Sc. Degree in Computer Science from Mahatma Gandhi University, Kottayam, Kerala and she is working as an Assistant Professor in the Department of Computer Science at KKTM Government College, Kodungallur, India. She is currently pursuing the Ph.D degree at Cochin University of Science and Technology, Cochin, India. Her research interests are in image processing, document image analysis, pattern recognition, multi-modal data analysis and handwriting recognition. She has published papers in international journals and conference proceedings.

Dr. Kannan Balakrishnan, born in 1960, received his M. Sc and M. Phil degrees in Mathematics from University of Kerala, India, M. Tech degree in Computer and Information Science from Cochin University of Science & Technology, Cochin, India and Ph. D in Futures Studies from University of Kerala, India in 1982, 1983, 1988 and 2006 respectively. He is currently working with Cochin University of Science & Technology, Cochin, India, as an Associate Professor in the Department of Computer Applications. He has visited Netherlands as part of a MHRD project on Computer Networks. Also he is the co investigator of Indo-Slovenian joint research project by Department of Science and Technology, Government of India. He has published several papers in international journals and national and international conference proceedings. His present areas of interest are Graph Algorithms, Intelligent systems, Image processing, CBIR and Machine Translation. He is a reviewer of American Mathematical Reviews and several other journals.

Dr. Pramod K. Vijayaraghavan, presently working as Associate Professor in the Department of Computer Applications, Cochin University of Science and Technology, Cochin, India. He has twenty years of teaching and research experience. He has published 25 papers in national and international journals. His areas of interest include Simulation and Modelling, Cryptography and Coding Theory, Mathematical Morphology, Pattern recognition and Data Mining.