Offline Handwritten Malayalam Character Recognition Based on Chain Code Histogram

Jomy John, Pramod K. V, Kannan Balakrishnan
Department of Computer Applications
Cochin University of Science and Technology
Kochi, Kerala, India
jomyeldos@gmail.com, pramod_k_v@cusat.ac.in, bkannan@cusat.ac.in

Abstract -- **Optical Character Recognition plays an important role in Digital Image Processing and Pattern Recognition. Even though ambient study had been performed on foreign languages like Chinese and Japanese, effort on Indian script is still immature. OCR in Malayalam language is more complex as it is enriched with largest number of characters among all Indian languages. The challenge of recognition of characters is even high in handwritten domain, due to the varying writing style of each individual. In this paper we propose a system for recognition of offline handwritten Malayalam vowels. The proposed method uses Chain code and Image Centroid for the purpose of extracting features and a two layer feed forward network with scaled conjugate gradient for classification.**

***Keywords*—Malayalam, Handwritten Character Recognition, Image Processing, Chain code, Feed forward network.**

I. INTRODUCTION

Recognition of handwritten characters has been a popular area of research for many years and still remains an open problem. It has a versatile range of application domain, including postal automation, bank check processing, automating of processing of large volumes of data, language based learning, ledgering catalogue for library and reading aid for blind. Even though ambient study on offline and online character recognition [1] had been performed on foreign languages like Chinese and Japanese, efforts on Indian script is still immature. The challenge that lies in the recognition of Indic script is mainly due to enormously large character set, varying writing style of each individual, high similarity between characters and distorted and broken characters. Hence extreme variation is observed in the collected samples. The proposed work is an attempt for offline handwritten character recognition (HCR) problem by concentrating mainly on chain code histogram and normalized chain code histogram features. The work is extended by adding centroid of the image as supplementary feature and it was found that the combination improves the result. The organization of the paper is as follows. A brief introduction to Existing OCR systems in Indian scripts is given in Section II. Section III describes peculiarities of Malayalam Script. In Section IV, implementation model is discussed. Section V covers preprocessing steps taken. Feature Extraction Method is dealt in Section VI. Classification method is explained VII. Results and Discussion is covered in Section VIII. Future work is discussed in Section IX and Section X concludes the paper.

II. EXISTING OCR TECHNIQUES IN INDIC SCRIPTS

An excellent review of OCR in Indic scripts is presented by U. Pal and B.B. Chaudhuri [2]. HCR system for Devnagari characters are proposed by S. Arora [3] based on shadow features, chain code histogram and intersection points of characters with a recognition accuracy of about 92.16%. A hybrid zone based feature extraction for recognition of four Indian numerical with nearest neighbor and support vector machine classifiers with a recognition accuracy of 97.85% is reported by S. V. Rajashekararadhya [4]. Another method for recognition of printed and handwritten mixed Kannada numerals is presented using multi-class SVM for recognition yielding a recognition accuracy of 97.76% [5].Only a few works are reported in Malayalam Handwritten Character Recognition (HCR). In Ref [6], Fuzzy-zoned normalized vector distance features are classified using class modular neural network considering 44 Malayalam characters. Accuracy reported was 78.87%. In another work [7], State space Point Distribution (SSPD) parameters derived from gray scale based SSM of handwritten character samples are utilized to obtain an accuracy of 73.03%. Remarkable works on the application of daubechie wavelet coefficients in HCR were reported by G. Raju [8] [9] [10] and Renju John [11]. Recognition based on intensity pattern of characters was proposed by Rahiman [12]. Bindu Philip [13] describes OCR for printed Malayalam characters using SVM.

III. MALAYALAM SCRIPT

Malayalam is one of the four major Dravidian languages of South India and one among the twenty two scheduled languages of India with official language status in the State of Kerala and Union territories of Lakshadweep and Mahe, spoken by around 3.5 crores of people and ranked eighth in terms of the number of speakers. Malayalam script is derived from the Grantha script, an inheritor of olden Brahmi script. It is in close propinquity 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. And many characters are distinct just with a small variation in appearance. It is syllabic in nature and alphabets are classified into vowels and consonants. Conjunct symbols are used to combine certain consonants.

978-1-4244-7926-9/11/$26.00 ©2011 IEEE

A. *Malayalam Character Set*

Malayalam language script consists of 15 vowels (Fig 1) and 36 consonants (Fig 2). Although Malayalam language has 10 numerals, ranging from 0 to 9, it is seldom in use. Instead Arabic numerals are used in practice. Even though Malayalam script has been standardized, people still used to write in both old script and new script. Some samples of segmented characters are displayed in Fig. 3.

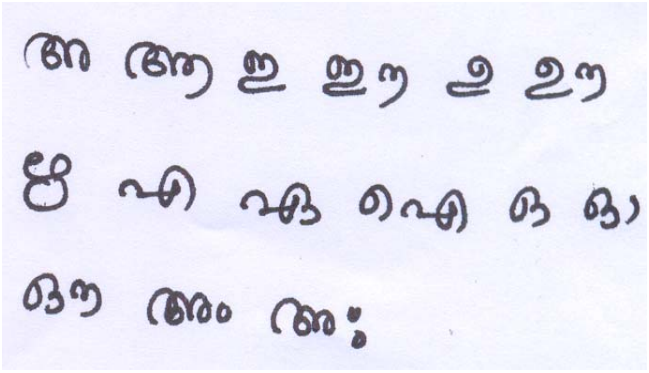

Figure 1. Handwritten vowels

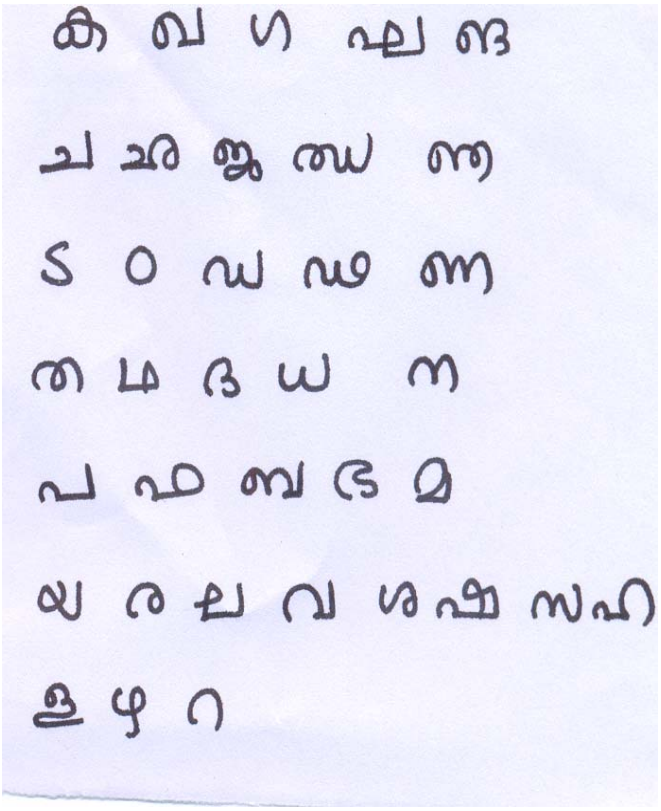

Figure 2. Handwritten consonants

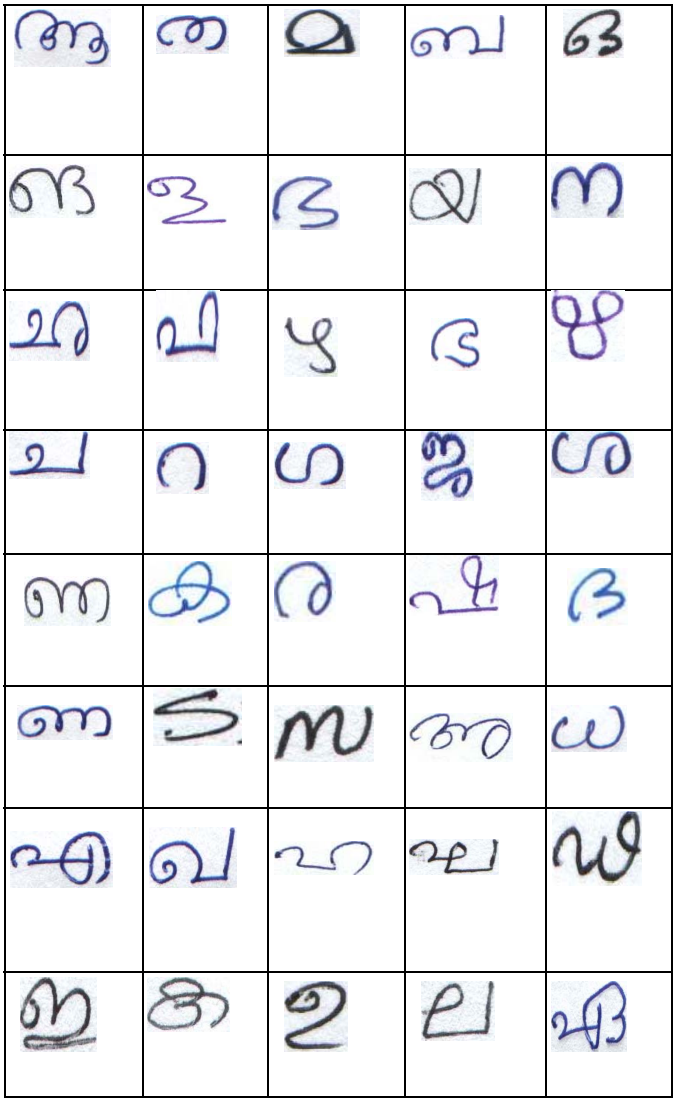

Figure 3. Some samples of segmented characters

IV. IMPLEMENTATION MODEL

The important steps of Character Recognition System include Preprocessing, Feature Extraction, Classification and Post Processing. Block diagram of a typical character recognition engine is shown in Fig. 4. The Preprocessing steps are depicted in Section V. Features are extracted based on statistical and structural features of images. Feature Extraction method used in this paper is described in Section VI. For classification Artificial Neural Networks and Support Vector Machines are used. Post processing includes error correction and mapping of characters into Unicode representation.

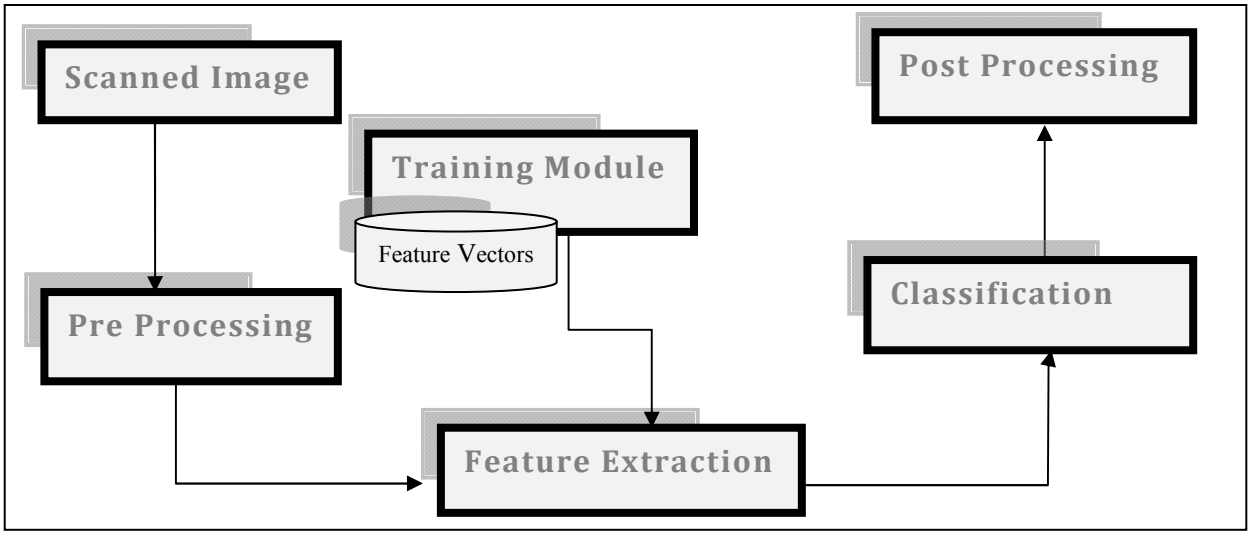

Figure 4. Block diagram of Character Recognition Engine

V. PREPROCESSING

Preprocessing is an essential step in Optical Character Recognition. The nature of preprocessing depends on subsequent steps. As a preliminary work, about 60 handwritten pages are collected from different persons containing characters in Malayalam language, without considering ink or pen variations. It contains broken and distorted characters also. Each page is scanned using 200, 300 or 600 DPI and stored either as BMP, JPG or TIF format. Each character is segmented using morphological method with a rectangular structuring element and the bounding box of each character image is stored BMP images. Each character is assigned a class id. Pre-processing steps used here are shown in Fig. 5. A median filter is applied to each segmented character image to reduce salt and pepper noise. The image is then converted to binary based on Otsu's [14] method of global image threshold. Edges in each binary image are found out. Image is filled with flood fill to avoid break in boundary contour. The results of all these processing are shown in Fig. 6. Character images are normalized to 256x256 using bicubic interpolation, where the output pixel value is a weighted average of pixels in the nearest 4 by 4 neighborhood. This size normalization avoids inter class variation among characters.

1. Noise Removal
2. Binarization
3. Segmentation
4. Size Normalization

Figure 5. Preprocessing steps

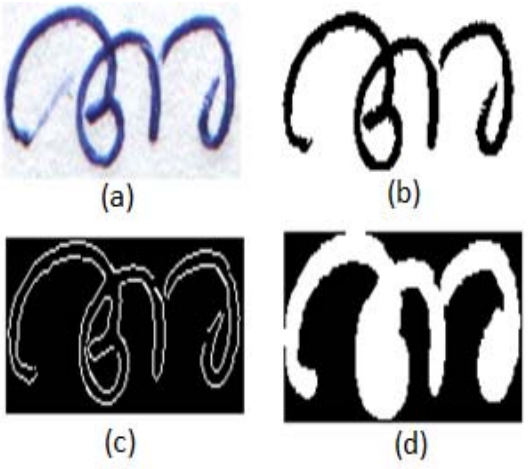

Figure 6. (a) Original Image (b) Binarized Image
(c) Edges of the image (d) Flood filled image

VI. FEATURE EXTRACTION

Feature extraction is a crucial step for any OCR system. There are several methods for the shape analysis of objects. Various feature extraction methods are covered in Trier.O.D [15]. In this paper, we are mainly concerned with the chain code based approach by Freeman [16]. Chain codes are used to represent the boundary by a connected sequence of straight line segments of specified length and direction [17]. It is an ordered sequence of *n* links

$$\{ x_i , i = 1 , \ldots , n \}$$

Where x_i is a vector connecting neighboring contour pixels. The directions of x_i are coded with integer values i=0,1,...,n-1.

A. *Chain code calculation of Handwritten Malayalam characters*

For extracting chain code features, edges of each size normalized segmented binary character image is traced, starting with the bottom most, left most point in the trajectory. The direction of each segment is coded both as four directional and as eight directional as in Fig. 7. The chain proceeds in clock wise manner and it is carried out till starting point is revisited. Then the chain code histogram (CCH) is calculated from the chain code representation of the contour. The CCH is a translation and scale invariant shape descriptor. To achieve better invariance the normalized chain code histogram (NCCH) is also used. Fig. 8 shows the plot of character image (ah), reconstructed from the boundary points.

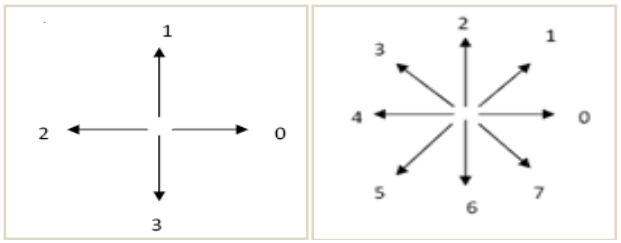

Figure 7. Directions of four connected and eight connected chain code

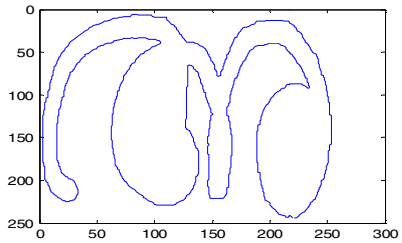

Figure 8. Plot of character image reconstructed from the boundary points

B. *Algorithm*

Apply the following algorithm for all character images on the database

Step 1: Resize character image to 256x256

Step 2: Binarize the input image

Step 3: Detect the edge

Step 4: Fill the character to avoid break in contour

Step 5: Extract the boundary points

Step 6: Obtain N directional chain code for
(i) N=4 (CCH4) and
(ii) N=8 (CCH8)

Step 7: Calculate the strength of values in N directions and normalize it, for
(i) N=4 (NCCH4) and
(ii) N=8 (NCCH8)

Step 8: Calculate the (x,y) points of the image centroid

Step 9: Construct input
Feature set I with 6(i) and 7(i)
Feature set II with 6(i), 7(i) and 8
Feature set III with 6(ii) and 7(ii)
Feature set IV with 6(ii), 7(ii) and 8

Repeat Step 1 to 9 for all images in the database

VII CLASSIFICATION

A two layer feed forward neural network as in Fig. 9 with sigmoid activation function is used for classification. The network is trained with scaled conjugate gradient (SCG) back propagation algorithm. This algorithm is based upon a class of optimization techniques in numerical analysis as the conjugate gradient methods using the second order information from

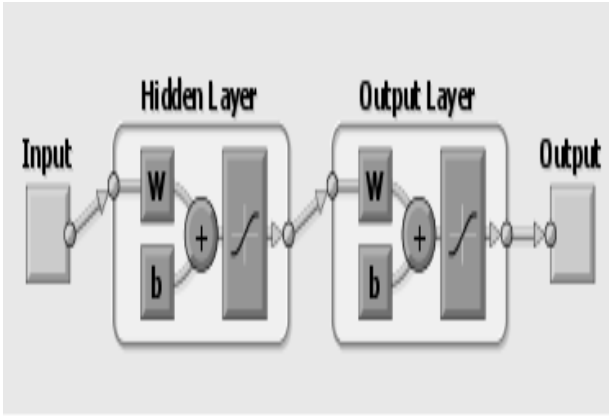

Figure 9. Neural Network Model

neural network, but requires only O(N) memory usage, where N is the number of weights in the network [18]. Mean squared error, which is the average squared difference between outputs and targets, is used as the performance measure.

VIII RESULTS AND DISCUSSION

In the first experiment 8 features as in Feature set I is given as input to the classifier. In the subsequent experiments, 10, 16 and 18 features as in Feature set II, III and IV, respectively, are used. From the set of samples 70% are used for training, 15% for validation and 15% for testing. The training, testing and validation samples are selected at random. Mean Squared Error is used as performance measure. The outcome is tabulated in table 1. The plot of training state and performance of feature set II are displayed in Fig. 10 and in Fig. 11 and that of feature set IV are displayed in Fig. 12 and Fig. 13 respectively.

TABLE 1: PERFORMANCE MEASURES

No	*Features used*	*Accuracy*			*Average Accuracy*
		Training	*Validation*	*Testing*	
I	4dir CCH and 4 dir NCCH	63.7	56.4	54.5	61.2%
II	4dir CCH, 4 dir NCCH and Centroid	68.8	61.8	63.6	66.9%
III	8 dir CCH and 8 dir NCCH	64.8	60.0	63.6	63.9%
IV	8dir CCH, 8dir NCCH and Centroid	72.3	72.7	70.9	72.1%

Tabulated Result

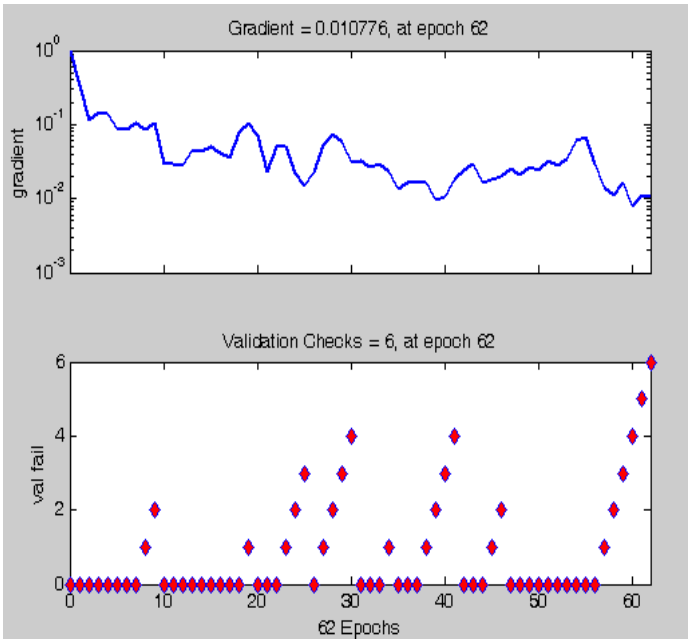

Figure 10. Training state of feature set II

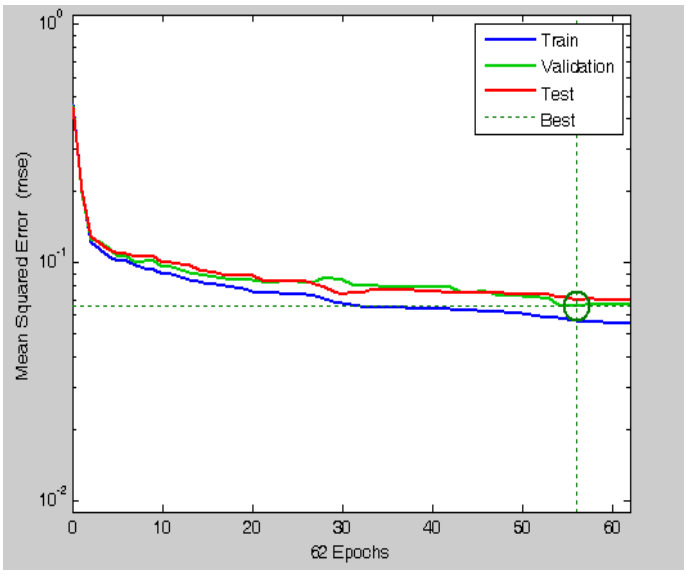

Figure 11. Performance plot with feature set II

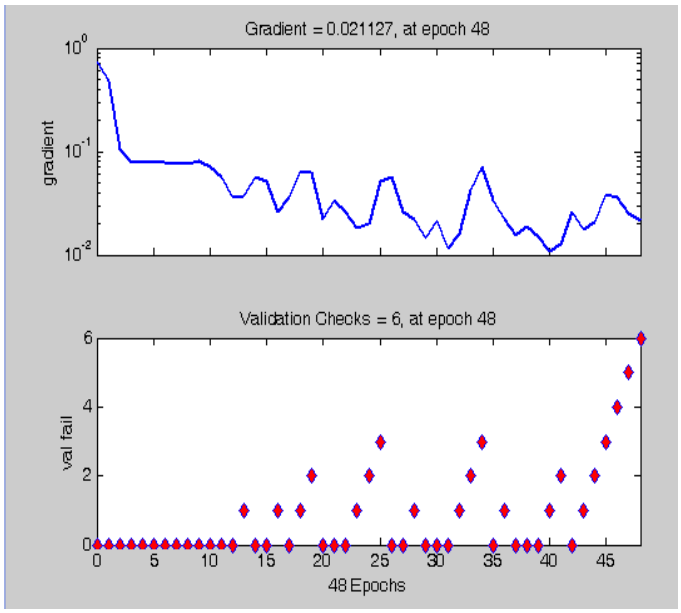

Figure 12. Training state of feature set IV

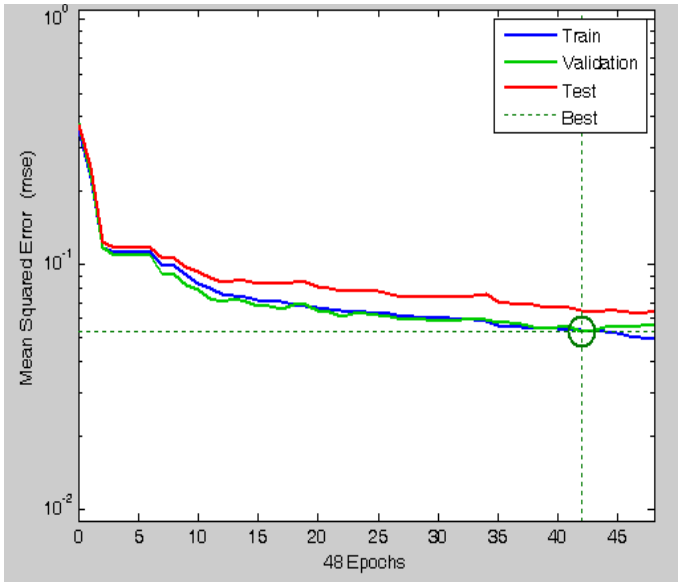

Figure 13. Performance plot with feature set IV

IX FUTURE WORK

A large collection of handwritten samples are a pre-requisite to the proper performance of any HCR. As a future plan, we would like to enhance the approach using sufficiently large number of samples and extend the work by using all characters in Malayalam language.

X CONCLUSION

A novel method for modeling Malayalam handwritten vowels based on both chain code histogram and normalized chain code histogram are introduced. Centroid of the image is used as an additional feature, which is found to improve the result.

REFERENCES

[1] R. Plamondan, S.N. Srihari, "Online and offline handwriting recognition: A comprehensive survey", IEEE Trans. On PAMI, Vol22(1) pp 63 – 84, 2000.

[2] U. Pal and B.B. Chaudhuri, "Indian script character recognition: A survey" ,Pattern Recognition, Elsevier ,Vol. 37, pp. 1887-1899, 2004.

[3] S. Arora, D. Bhattacharjee, M. Nasipuri, D K. Basu and M. Kundu, "Combining multiple feature extraction techniques for handwritten Devnagari character recognition " , 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA, December 8-10.

[4] S.V. Rajashekararadhya, Vanaja Ranjan P, "Zone-based hydrid feature extraction algorithm for handwritten numeral recognition of four Indian scripts",Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, San Anonio, TX, USA- October 2009.

[5] G. G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, "Printed and handwritten mixed Kannada numerals recognition using SVM", International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1622-1626

[6] Lajish V. L., "Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks", Proc. 4th Int. National conf. on Innovations in IT, 2007, 188 – 192.

[7] Lajish V. L., "Handwritten character recognition using gray scale based state space parameters and class modular NN",Proc. 4th Int. National conf. on Innovations in IT, 2007, 374 – 379.

[8] G. Raju, "Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients", Proc. of 14th International conference on Advanced Computing and Communications, 2006, pp 217 – 221

[9] G. Raju, "Wavelet transform and projection profiles in handwritten character recognition- A performance analysis", Proc. Of 16th International Conference on Advanced Computing and Communications, Chennai 2008, pp309-314.

[10] G. Raju and K. Revathy, "Wavepackets in the recognition of isolated handwritten characters", Proceedings of the World Congress on Engineering 2007 Vol IWCE 2007, July 2 - 4, 2007, London, U.K.

[11] Renju John, G. Raju and D. S. Guru, "1D Wavelet transform of projection profiles for isolated handwritten character recognition", Proc. of ICCIMA07, Sivakasi, 2007, 481-485, Dec 13-15.

[12] M Abdul Rahiman et. al., "Isolated handwritten Malayalam character recognition using HLH intensity patterns", 2010 Second International Conference on Machine Learning and Computing

[13] Bindu Philip, R.D. Sudhakaer Samuel, "Preferred computational approaches for the recognition of different classes of printed Malayalam characters using hierarchical SVM classifiers", International Journal of Computer Applications (0975-8887) vol 1-No.16,2010

[14] Otsu.N, "A threshold selection method from gray level histograms", IEEE Trans. Systems, Man and Cybernetics, vol.9, pp.62-66, 1979

[15] Trier.O.D, Jain.A.K and Taxt.J, "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662, 1996.

[16] Freeman, H., On the encoding of arbitrary geometric configurations IRE Trans. on Electr. Comp. or TC(10), No. 2, June, 1961, pp. 260-268

[17] Rafael C. Gonzalez, Richard E. Woods,"Digital Image Processing (2nd Edition) ",PHI

[18] Martin Fodslette Møller, "A scaled conjugate gradient algorithm for fast supervised learning", Neural Networks, Elsevier, Volume 6, Issue 4, 1993, pp 525-533