

A Comparative Study of Different Feature Extraction Techniques for Offline Malayalam Character Recognition

Anitha Mary M.O. Chacko and P.M. Dhanya

Abstract Offline Handwritten Character Recognition of Malayalam scripts have gained remarkable attention in the past few years. The complicated writing style of Malayalam characters with loops and curves make the recognition process highly challenging. This paper presents a comparative study of Malayalam character recognition using 4 different feature sets—Zonal features, Projection histograms, Chain code histograms and Histogram of Oriented Gradients. The performance of these features for isolated Malayalam vowels and 5 consonants are evaluated in this study using feedforward neural networks as classifier. The final recognition results were computed using a 5 fold cross validation scheme. The best recognition accuracy of 94.23 % was obtained in this study using Histogram of Oriented Gradients features.

Keywords Offline character recognition • Feature extraction • Neural networks

1 Introduction

Offline character recognition is the process of translating handwritten text from scanned, digitized or photographed images into a machine editable format. Compared to online recognition, offline recognition is a much more challenging task due the lack of temporal and spatial information. Character recognition research has gained immense popularity because of its potential applications in the areas of postal automation, bank check processing, number plate recognition etc. Even though ambient studies have been performed in foreign languages [1], only very

A.M.M.O. Chacko (✉) · P.M. Dhanya
Department of Computer Science and Engineering, Rajagiri School of Engineering
and Technology, Kochi, India
e-mail: anithamarychacko@gmail.com

P.M. Dhanya
e-mail: dhanya_pm@rajagiritech.ac.in

few works exist in the Malayalam character recognition domain. This is mainly due to its extremely large character set and complicated writing style with loops curves and holes.

Some of the major works reported in the Malayalam character recognition domain are as follows: Lajish [2] proposed the first work in Malayalam OCR using fuzzy zoning and normalized vector distances. 1D wavelet transform of vertical and horizontal projection profiles were used in [3] for the recognition of Malayalam characters. The performance of wavelet transform of projection profiles using 12 different wavelet filters were analyzed in [4]. In [5], recognition of Malayalam vowels was done using chain code histogram and image centroid. They have also proposed another method for Malayalam character recognition using Haar wavelet transform and SVM classifier [6]. Moni and Raju used Modified Quadratic Classifier and 12 directional gradient features for handwritten Malayalam character recognition [7]. Here gradient directions were computed using Sobel operators and were mapped into 12 directional codes. Recently, Jomy John proposed another approach for offline Malayalam recognition using gradient and curvature calculation and dimensionality reduction using Principal Component Analysis (PCA) [8]. A detailed survey on Malayalam character recognition is presented in [9].

A general handwritten character recognition system consists of mainly 4 phases—Preprocessing, Feature Extraction, Classification and Postprocessing. Among these, feature extraction is an important phase that determines the recognition performance of the system. To get an idea of recognition results of different feature extraction techniques in Malayalam character recognition, we have performed a comparative study using 4 different features—Zonal features, projection histograms, chain codes and Histogram of Oriented Gradients (HOG) features. The performance of these four feature sets are analyzed by using a two layer feedforward neural network as classifier.

The paper is structured as follows: Sect. 2 presents the data collection method used and the sequence of preprocessing steps done. Section 3 describes the feature extraction procedure for the four feature sets. The classifier used is introduced in Sect. 4. Section 5 presents the experimental results and discussions and finally conclusion is presented in Sect. 6.

2 Data Collection and Preprocessing

Malayalam belongs to the Dravidian family of languages which has official language status in Kerala. The complete character set of Malayalam consists of 15 vowels, 36 consonants, 5 chillu, 9 vowel signs, 3 consonant signs, 3 special characters and 57 conjunct consonants. Since a benchmarking database is not available for Malayalam, we have created a database of 260 samples for the isolated Malayalam vowels and 5 consonants ('ka', 'cha', 'tta', 'tha' and 'pa'). For this 13 class recognition problem, we have collected handwritten samples from 20 people

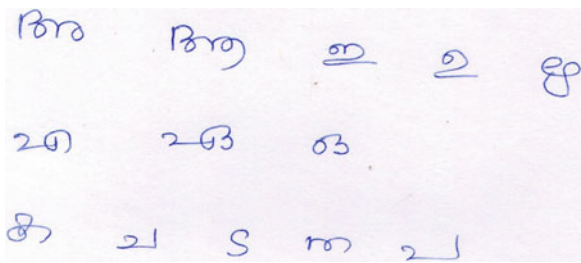


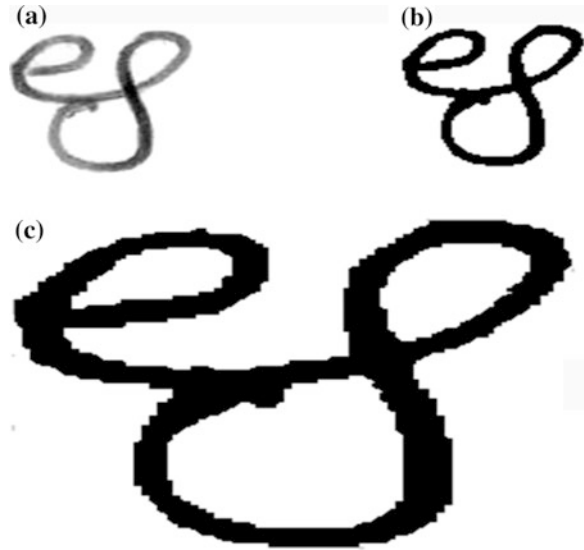
Fig. 1 Samples of handwritten Malayalam characters

belonging to different age groups and professions. Each of these 13 characters are assigned class-ids. The scanned images were then subjected to preprocessing. Figure 1 shows sample characters of the database.

2.1 Preprocessing

Preprocessing steps are carried out to reduce variations in the writing style of different people. The sequences of preprocessing steps (Fig. 2) carried out are as follows: Here, scanned images are binarized using Otsu’s method of global thresholding. This method is based on finding the threshold that minimizes the intra-class variance. A large amount of noise such as salt and pepper noise may exist in the image acquired by scanning. So in order to reduce this noise to some extent, we have applied a 3×3 median filter. In the segmentation process, the

Fig. 2 Preprocessing steps.
a Scanned image. **b** Binarized image. **c** Size normalized image



character images are separated into individual text lines from which characters are isolated using connected component labeling. Finally, the images are resized to 256×256 using bicubic interpolation techniques. This operation ensures that all characters have a predefined height and width.

3 Feature Extraction

The performance of an HCR system depends to a great extent on the extracted features. Over the years, many feature extraction techniques have been proposed for character recognition. A survey of feature extraction techniques is presented in [10]. In this study, we have used 4 sets of features for comparing the performance of the character recognition system: Zonal features, Projection histograms, Chain code histograms and Histogram of Oriented Gradients.

3.1 Zoning

Zoning is a popular method used in character recognition tasks. In this method, the character images are divided into zones of predefined sizes and then features are computed for each of these zones. Zoning obtains local characteristics of an image. Here, we have divided the preprocessed character images into 16 zones (4×4) as in and then pixel density features were computed for each of the zones (Fig. 3). The average pixel density was calculated by dividing the number of foreground pixels by the total number of pixels in each zone i .

$$d(i) = \frac{\text{Number of foreground pixels in zone } i}{\text{Total number of pixels in zone } i} \quad (1)$$

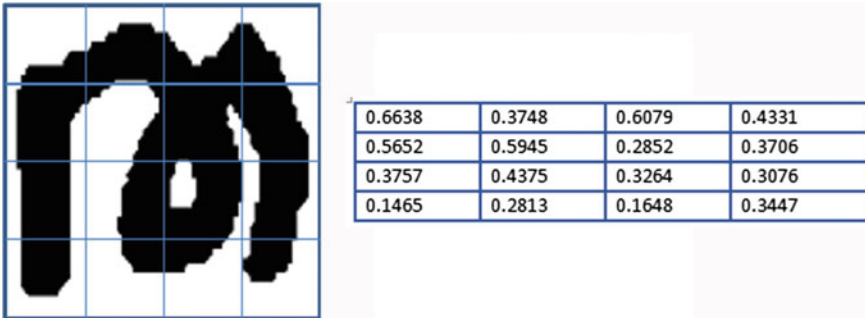


Fig. 3 4×4 zoning

Thus we have obtained 16 density features which are used as input to the classifier.

3.2 Projection Profile

Projection profile is an accumulation of black pixels along rows or columns of an image. The discriminating power of horizontal and vertical projection profiles make them well suitable for the recognition of a complex language like Malayalam. Projection profiles have been successfully applied for Malayalam character recognition [3, 4].

In this study, we have extracted both vertical and horizontal projection profiles by counting the pixels column wise and row wise respectively which together forms a 512 dimension feature vector (Fig. 4 shows the vertical and horizontal projection histogram for a Malayalam character ‘tha’).

Since, the size of the feature vector is too large, we have applied Principal Component Analysis (PCA) to reduce the dimensionality of the feature set. PCA is a technique that reduces the dimensionality of the data while retaining as much variations as possible in the original dataset. Using PCA, we have reduced the dimension of the feature vector from 512 to 260.

3.3 Chain Code Features

The chain code approach proposed by Freeman [11] is a compact way to represent the contour of an object. The chain codes are computed by moving along the boundary of the character in clockwise/anticlockwise direction and assigning each

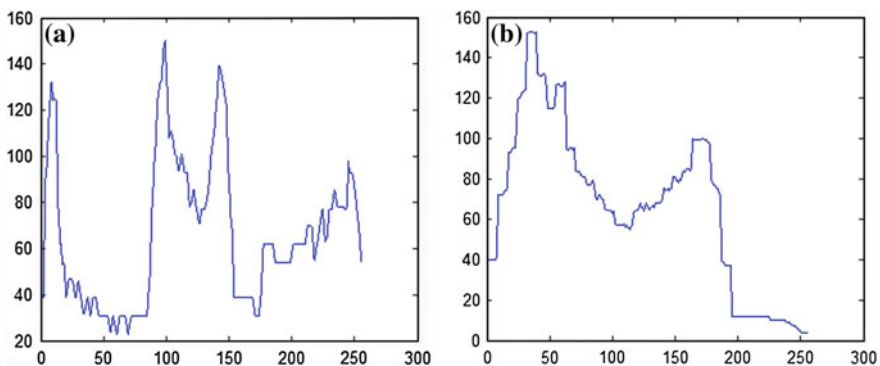
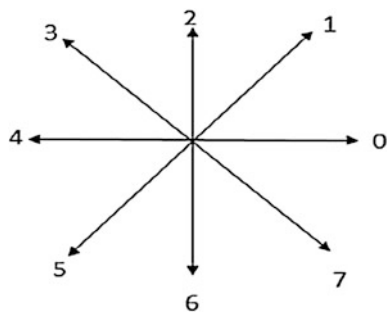


Fig. 4 Projection histogram of character ‘tha’. **a** Horizontal projection histogram. **b** Vertical projection histogram

Fig. 5 8 directional chain codes



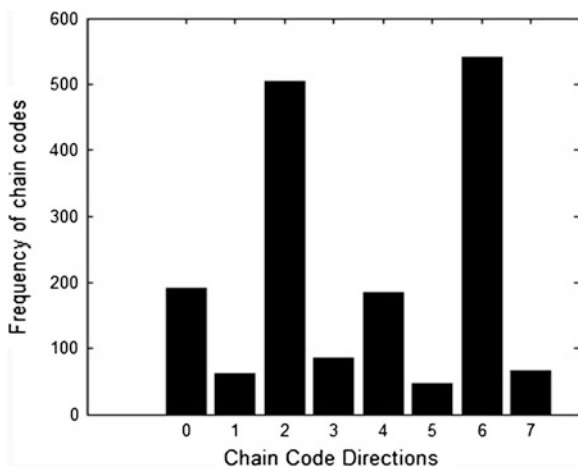
pixel along the contour a direction code (0–7) that indicates the direction of next pixel along the contour till the starting point is revisited. Here, we have used freeman chain code of eight directions (Fig. 5). Chain code and image centroid have been successfully applied for Malayalam vowel recognition in [5, 14].

Since the size of the chain code varies for different characters, we normalize it as follows: The frequency of each direction code is computed to form a chain code histogram (CCH). Figure 6 shows the chain code histogram of Malayalam character ‘tha’. Image centroid is also used as an additional feature here. Thus we get a feature vector of size 10.

3.4 Histogram of Oriented Gradients

Histograms of Oriented Gradients are feature descriptors that are computed by counting the occurrences of gradient orientations in localized parts of an image. For computing these features, the image is divided into cells and histograms of gradient

Fig. 6 Chain code histogram



directions are formed for each of these cells. These histogram forms the descriptor. HOG features have been successfully implemented for other applications such as human detection [12], pedestrian detection [13] etc. Recently, it has also been implemented for character recognition in Hindi. However these features have not been explored for Malayalam character recognition.

In this method, the image was divided into 9 overlapping rectangular cells and for each of these cells, gradient directions were computed. Based on the gradient directions, each pixel within a cell casts a weighted vote to form an orientation based histogram channel of 9 bins. The gradient strength of each cell were normalized according to L2-norm. Thus the 9 histograms with 9 bins are concatenated to form an 81 dimensional feature vector [13] which is fed as input to the classifier.

4 Classification

Classification is the final stage of character recognition task in which character images are assigned unique labels based on the extracted features. In this study, we have used neural networks for comparing the performance of different feature sets. The principal advantage of neural networks is that they can learn automatically from examples. Here, we have used a two layer feedforward neural network consisting of a single hidden layer (Fig. 7). The input to the neural network consists of each of the feature sets that we have extracted. Thus the number of nodes in the input layer is equal to the size of the feature set that we use in each case. The output layer contains one node for each of the output classes, i.e., here we have 13 nodes in the output layer. We have used the number of hidden layer neurons to be 20 for our experiment.

5 Experimental Results and Discussions

In this section, the results of different feature sets for offline Malayalam character recognition are presented. The implementation of the system was carried out using Matlab R2013a. The results have been computed using a 5 fold cross validation

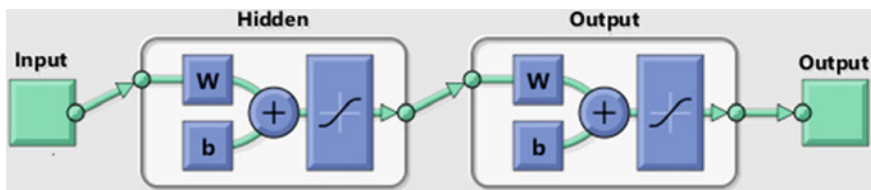


Fig. 7 Neural network model

technique. In a five-fold cross validation scheme, the entire dataset is divided into 5 subsets. During each fold, one of the subset is used for testing the classifier and the rest are used for training. The recognition rates from the test set in each fold are averaged to obtain the final accuracy of the classifier.

Table 1 summarizes the overall accuracy of the system with each of the four feature sets. A graphical representation of the recognition results is also shown (Fig. 8). From the experiment, the best recognition rate of 94.23 % was obtained for the histogram of oriented gradients features. The next highest accuracy was obtained for projection histograms with PCA. But compared to other feature sets, they need feature vectors of larger size (260). The density based features also provide good recognition accuracy with a relatively small feature vector size of 16. The chain code histogram feature gave the lowest recognition results among all the features used in this study. The recognition rate of 78.8 % was obtained using this chain code histogram feature.

From the confusion matrix obtained for each of the feature sets, we calculate precision and recall values. The plot of precision and recall values for each of the four feature sets are shown in Figs. 9 and 10 respectively. From the graphs, we noted that the HOG features achieved the highest average precision and recall values of 0.9423 and 0.9449 respectively. The lowest precision and recall values of 0.8026 and 0.7924 were shown by chain code histogram features. The average

Table 1 Recognition results of different feature sets

Feature set	Feature size	Accuracy (%)
Zonal density	16	84.6
Projection histograms + PCA	260	88.07
Chaincode histogram + centroid	10	78.8
HOG	81	94.23

Fig. 8 Recognition accuracy of different features

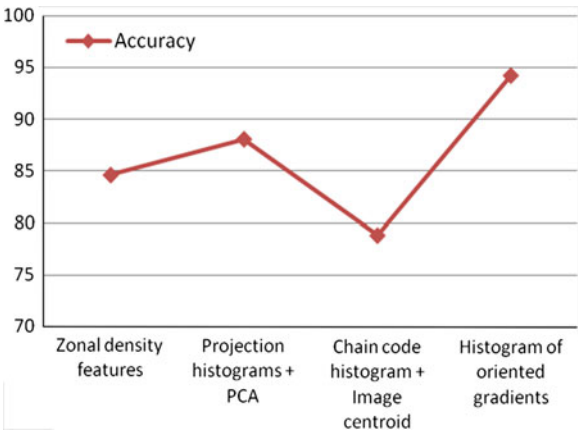


Fig. 9 Precision versus classes

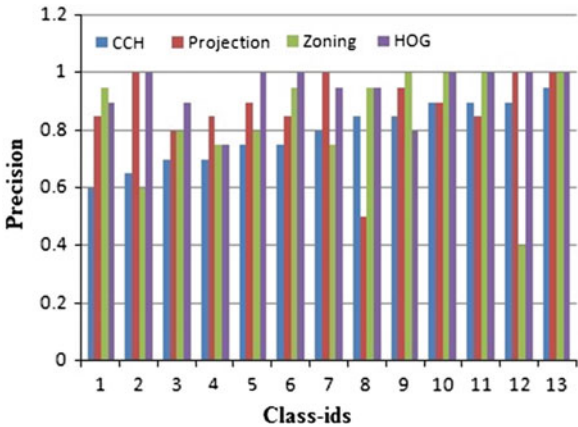
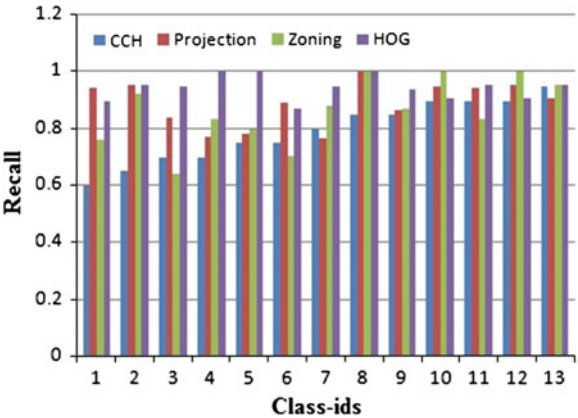


Fig. 10 Recall versus classes



precision values for the zonal density features and projection features were 0.8423 and 0.8808 respectively. The average recall values for these features were 0.8614 and 0.8904 respectively.

6 Conclusion

The work presented in this paper analyses the performance of offline Malayalam character recognition using 4 feature sets—zone density features, projection histograms, chain code histograms and histogram of oriented gradient features. These features were classified using a two layer feedforward neural network. A five fold cross validation scheme was applied to measure the performance of the system. We have obtained the best recognition accuracy of 94.23 % using histogram of oriented gradient features. This accuracy can be further improved by using a larger dataset

for training. Also post processing approaches for identifying similar shaped characters can further improve the recognition rate. The authors hope that this paper aids researchers who are working on Malayalam character recognition domain in their future works.

References

1. Plamondon, R., Srihari, S.N.: Online and offline character recognition: a comprehensive survey. *IEEE Trans. PAMI* **22**, 63–84 (2000)
2. Lajish, V.L.: Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks. In: *Proceedings of 4th International National Conference on Innovations in IT*, pp. 188–192 (2007)
3. John, R., Raju, G., Guru, D.S.: 1D wavelet transform of projection profiles for isolated handwritten character recognition. In: *Proceedings of ICCIMA07*, pp. 481–485, Sivakasi (2007)
4. Raju, G.: Wavelet transform and projection profiles in handwritten character recognition—a performance analysis. In: *Proceedings of 16th International Conference on Advanced Computing and Communications*, pp. 309–314, Chennai (2008)
5. John, J., Pramod K.V., Balakrishnan K.: Offline handwritten Malayalam character recognition based on chain code histogram. In: *Proceedings of ICETECT* (2011)
6. John, J., Pramod, K.V., Balakrishnan, K.: Unconstrained handwritten Malayalam character recognition using wavelet transform and support vector machine classifier. In: *International Conference on Communication Technology and System Design*, Elsevier (2011)
7. Moni, B.S., Raju, G.: Modified quadratic classifier and directional features for handwritten Malayalam character recognition. In: *IJCA Special Issue on Computational Science—New Dimensions Perspectives NCCSE* (2011)
8. John, J., Balakrishnan, K., Pramod, K.V.: A system for offline recognition of handwritten characters in Malayalam script. *Int. J. Image Graph. Signal Process.* **4**, 53–59 (2013)
9. Chacko, A.M.M.O.: Dhanya PM, Handwritten character recognition in Malayalam scripts—a review. *Int. J. Artif. Intell. Appl. (IJAA)* **5**(1), 79–89 (2014)
10. Trier, O.D., Jain, A.K., Taxt, J.: Feature extraction methods for character recognition—a survey. *Pattern Recogn.* **29**(4), 641–662 (1996)
11. Freeman, H.: On the encoding of arbitrary geometric configurations. *IRE Trans. Electr. Comp. TC* **10**(2), 260–268 (1961)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886–893 (2005)
13. Ludwig, O., Delgado, D., Goncalves, V., Nunes, U.: Trainable classifier-fusion schemes: an application to pedestrian detection. In: *12th International IEEE Conference on Intelligent Transport Systems*, pp. 1–6 (2009)
14. Chacko, A.M.M.O., Dhanya, P.M.: A differential chain code histogram based approach for offline Malayalam character recognition. In: *International Conference on Communication and Computing (ICC-2014)*, pp. 134–139 (2014)

Computational Intelligence in Data Mining - Volume 2
Proceedings of the International Conference on CIDM,
20-21 December 2014

Jain, L.C.; Behera, H.S.; Mandal, J.K.; Mohapatra, D.P.
(Eds.)

2015, XXVIII, 707 p. 276 illus., Hardcover

ISBN: 978-81-322-2207-1