Combining Classifiers for Offline Malayalam Character Recognition

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

Department of Computer Science & Engineering
Rajagiri School of Engineering & Technology
Kochi, India
anithamarychacko@gmail.com, dhanya_pm@rajagiritech.ac.in

Abstract. Offline Character Recognition is one of the most challenging areas in the domain of pattern recognition and computer vision. Here we propose a novel method for Offline Malayalam Character Recognition using multiple classifier combination technique. From the preprocessed character images, we have extracted two features: Chain Code Histogram and Fourier Descriptors. These features are fed as input to two feedforward neural networks. Finally, the results of both neural networks are combined using a weighted majority technique. The proposed system is tested using two schemes- Writer independant and Writer dependant schemes. It is observed that the system achieves an accuracy of 92.84% and 96.24% respectively for the writer independant and writer dependant scheme considering top 3 choices.

Keywords: Character Recognition, Fourier Descriptors, Chain Code Histogram, Multiple Classifier Combination, Neural Networks.

1 Introduction

Offline character recognition is the task of recognizing handwritten text from a scanned, digitized or photographed sheet of paper and converting it to a machine editable format. It is an active research field in the domain of pattern recognition and machine vision due to the intrinsic challenges present in them. Further, the large variations in the writing style among different writers and even among the same writers at different times complicate the recognition process. The numerous applications of handwritten character recognition in the domain of postal automation, license plate recognition and preservation of historical and degraded documents makes it an interesting research area.

OCR research is at a well advanced stage in foreign languages like English, Chinese, Latin and Japanese [1]. However, compared to these languages, OCR research in Indic scripts is far behind. The highly complicated and similar writing style among different characters, the large character set and the existence of old and new scripts are some of the factors that attribute to the highly challenging nature of the recognition process in these scripts. Though many works have been reported in the past few years, a complete OCR system for Indian

S.C. Satapathy et al. (eds.), *Emerging ICT for Bridging the Future – Volume 2*,
Advances in Intelligent Systems and Computing 338, DOI: 10.1007/978-3-319-13731-5_3

languages is still lacking. Among the Indian languages, the research on South Indian languages such as Kannada, Tamil, Telugu and Malayalam demands far more attention.

An excellent survey on OCR research in South Indian Scripts is presented in [2],[3]. A detailed survey on handwritten malayalam character recognition is presented in [4]. Rajashekararadhya et al. [5] proposed an offline handwritten numeral recognition system for the four South Indian languages- Malayalam, Tamil, Kannada and Telugu. The proposed system used zone centroid and image centroid based features. Here, Nearest Neighbour and Feedforward neural networks were used as classifiers. The first work in Malayalam OCR was reported by Lajish V.L. [6] using fuzzy zoning and normalized vector distances for the recognition of 44 isolated basic Malayalam characters. A character recognition scheme using Run length count (RLC) and MQDF classifier was proposed in [7]. In [8], a character recognition technique using cross features, fuzzy depth features, distance features and Zernike moments and a Probabilistic Simplified Fuzzy ARTMAP classifier was proposed.

The above HCR systems have proposed many feature extraction techniques and classifiers but a remarkable achievement has not yet been obtained in Malayalam character recognition. All the above works have been based on a single classifier scheme. Even though multiple classifier schemes have been applied to other Indic languages [9], its effect on a complicated language like Malayalam has not been explored till now. This motivated us to investigate the effects of using a multiple classifier system for the recognition of handwritten malayalam characters.

In this paper, we propose a multiple classifier system for the recognition of handwritten Malayalam characters. The extracted features are based on the chain code histogram and fourier descriptors. These features are fed as input to two feedforward neural networks. The final results are obtained by combining the results of individual classifiers using a weighted majority scheme. To the best of our knowledge, the use of a multiple classifier system for the recognition of malayalam characters represents a novelty.

This paper is structured as follows: Section 2 introduces the architecture of the proposed system. Section 3 presents the experimental results and finally, Section 4 concludes the paper.

2 Proposed System

The proposed system consists of mainly 4 stages: Preprocessing, Feature Extraction, Classification and Post Processing. The scanned image is first subjected to preprocessing to remove as much distortions as possible. After preprocessing, chain code histogram features and fourier descriptors are extracted and fed as input to two neural network classifiers. The output of these networks are combined using a weighted majority scheme to obtain the final recognition results. Finally, the characters are mapped to their corresponding Unicode values in the post processing stage. Fig.1 shows the architecture of the proposed system.

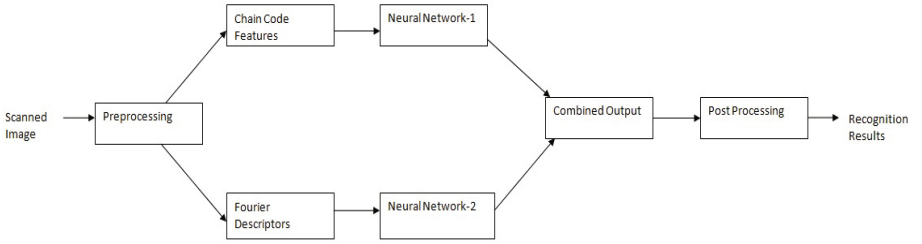

Fig. 1. Architecture of the Proposed System

2.1 Preprocessing

The objective of preprocessing phase is to eliminate as much distortions as possible from the scanned image. These distortions occur due to the poor quality of scanners and degraded documents. A 3x3 median filter is applied to remove salt and pepper noise to a certain extent. The scanned image is converted to binary using Otsu's method of global thresholding. This approach separates the foreground and background pixels. Then the line segmentation and character segmentation of the collected samples are carried out using horizontal and vertical projection profiles. Fig.2 shows the segmentation operation using these methods. Further, the characters are cropped by placing a bounding box around it. These cropped characters are finally normalized to 36x36 using bilinear transformation.

2.2 Feature Extraction

Feature extraction is the process of extracting relevant features from character images which are used as input to the classifier. This is an important phase in character recognition as effective features contribute to the success rate of the classifier. Here, we have used two different feature sets based on chain code histogram and invariant fourier descriptors. These features are obtained after the boundary of the image is extracted.

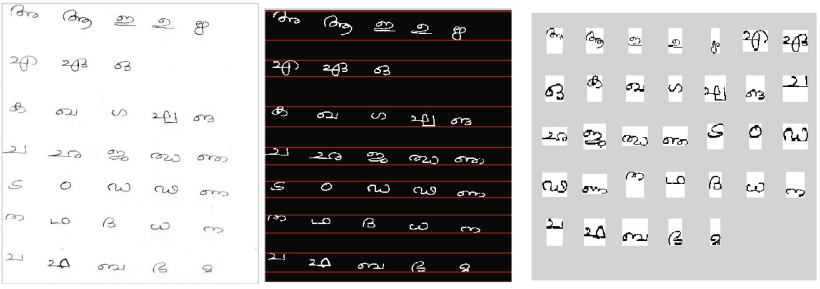

Fig. 2. Line and Character Segmentation

2.2.1 Chain Code Histogram

C hain code is a compact way to represent the boundary of an image. In the eight directional chain code approach proposed by Freeman [10] , the direction of each pixels are coded with integer values of 0-7. Here the chain codes are obtained by moving along the boundary in clockwise direction. The chain code histogram is obtained by computing the frequency of each direction chain codes. These chain code histograms are then normalized by dividing the frequencies of each direction by the total number of chain code values. Thus the input to the first neural network consists of 16 features.

2.2.2 Fourier Descriptors

T he fourier descriptors are powerful boundary based descriptors that represents the shape in the frequency domain. The advantage of fourier descriptors is that the overall shape features can be approximated using only a fixed number of discrete Fourier coefficients. Suppose the set of coordinates(x_k ,y_k) describing the contour of a shape consists of N pixels from 0 to N-1. The two dimensional plane can be interpreted as a complex plane where the horizontal axis represents the real part and the vertical axis represents the imaginary part. Thus, the contour can be represented as a sequence of complex numbers of the form:

$$z(k) = x(k) + iy(k) \tag{1}$$

Thus this representation reduces a 2D problem into a 1D problem. The discrete fourier transform of z(k) is:

$$a(u) = \sum_{k=0}^{N-1} z(k) e^{\frac{-j2\pi uk}{N}} \quad u = 0..N-1 \tag{2}$$

The complex coefficients a(u) are called the Fourier descriptors of the boundary. Here, translation invariance is obtained by discarding the zeroth fourier descriptor as it is the only one that contributes to translation. Scale invariance is achieved by dividing all descriptors by the absolute value of the first descriptor. Rotation and shift invariance is achieved by discarding the phase information and using only the absolute values of the fourier descriptors. We have chosen the number of descriptors to be 10 for our experiment.

2.3 Classification

In this stage, characters are mapped to unique labels based on the features extracted. We have used two feedforward neural networks for the two feature sets consisting of the chain code features and fourier descriptors respectively. Each of the network were trained with Bayesian regulation backpropagation algorithm. The classification for the individual classifiers were obtained by the maximum response strategy. The results of both the individual classifiers were combined using a weighted majority scheme to obtain the final recognition results.

2.3.1 Classifier Combination

C ombining individual classifiers overcomes the limitations of single classifiers in solving difficult pattern recognition problems involving large number of output classes. A proper combination of multiple classifiers provide more accurate recognition results than individual classifiers since each classifier offers complementary information about the pattern to be classified. We have used two feedforward neural networks trained with 16 chain code features and 10 fourier descriptors respectively. The weighted majority scheme[9] used to combine results of these neural networks are as follows:The final combined decision d_{com} is:

$$d_{com} = max \sum_{k=1}^{N} w_k * \delta_{ik} \quad 1 \leq i \leq N_c \tag{3}$$

where

$$w_k = \frac{d_k}{\sum_{k=1}^{N} d_k} \tag{4}$$

Here δ_{ik} denotes the k^{th} classifier decision to assign an unknown pattern to class i, N indicates the number of classifiers , N_c indicates the number of output classes, w_k denotes the weight assigned to k^{th} classifier and $d_k$indicates the recognition rate of k^{th} classifier

3 Experimental Results

The experiment was conducted on 33 characters-8 isolated vowels and 25 consonants of Malayalam character set. A standard benchmark database is not available for Malayalam. Fig. 3 shows the sample characters that we have used for our study and the Class-ids assigned to each of them. We have tested the system according to two different schemes: Writer dependant(Scheme 1) and writer independent schemes(Scheme 2). For the writer independent scheme, the database consists of 825 samples of handwritten data collected from 25 people belonging to different age groups and professions and for the writer dependant scheme, the database consists of another 825 samples collected from 5 different people. In the writer dependant scheme, writing samples of people which were used in training were subjected to testing whereas in the writer independent scheme, the writing samples of new users whose samples were not used for training were subjected to testing.

In both schemes, 80% of the samples were used for training and 20% were used for testing. The results are summarized in Fig. 4 and Table 1. The recognition accuracy obtained from the chain code histogram classifier and fourier descriptor classifier are 80.61% and 66.67% respectively for the writer independant scheme.

Class-Id	Character	Class-Id	Chracter	Class-Id	Character	Class-Id	Character
1	അ	2	ആ	3	ഇ	4	ഉ
5	ഋ	6	എ	7	ഏ	8	ഒ
9	ക	10	ഖ	11	ഗ	12	ഘ
13	ങ	14	ച	15	ഛ	16	ജ
17	ഝ	18	ഞ	19	ട	20	ഠ
21	ഡ	22	ഢ	23	ണ	24	ത
25	ഥ	26	ദ	27	ധ	28	ന
29	പ	30	ഫ	31	ബ	32	ഭ
33	മ						

Fig. 3. Database Samples

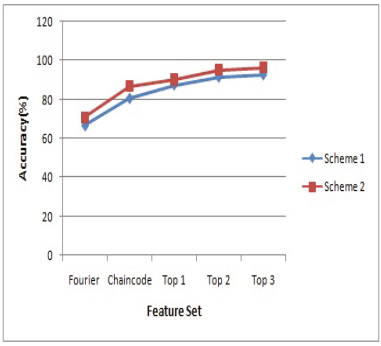

Fig. 4. Performance of Scheme 1 and Scheme 2

Table 1. Experimental Results

Feature Set(%)	Accuracy(%)	
	Scheme 1	Scheme 2
Chain Code	80.61	86.79
Fourier Descriptor	66.67	71.03
Combined(Top 1)	87.63	90.18
Combined(Top 2)	91.63	95.27
Combined(Top 3)	92.84	96.24

The combined classifier gives an overall accuracy of 92.84% as we considered top 3 choices and 87.63% for the top 1 choices for the entire database samples.

Compared to writer independent scheme, a higher recognition accuracy was obtained for the writer dependant scheme. For the writer dependant scheme, a recognition accuracy of 86.79% and 71.03% were obtained for the chain code histogram classifier and fourier descriptor classifier respectively. The combined classifier gives an overall accuracy of 96.24% as we considered top 3 choices and 90.18% for the top 1 choices. Based on the confusion matrix of top 1 choice obtained for both schemes, we have calculated seven useful measures: Precision, Recall(TP Rate/Sensitivity), Specificity(TN Rate), FP Rate and F-score. The results of both schemes are summarized in Table 2.

Table 2. Classification Results: Scheme 1 and Scheme 2

Class-Id	Writer Independant Scheme					Writer Dependant Scheme				
	Precision	Recall	Specificity	FP-Rate	F-score	Precision	Recall	Specificity	FP-Rate	F-score
1	0.6800	0.8947	0.9901	0.0099	0.7727	0.8800	0.9565	0.9963	0.0037	0.9167
2	0.9600	0.9231	0.9987	0.0013	0.9412	0.9600	0.8889	0.9987	0.0013	0.9231
3	0.9600	0.8889	0.9987	0.0013	0.9231	0.8400	0.9130	0.9950	0.0050	0.8750
4	0.8800	0.9565	0.9963	0.0037	0.9167	0.9200	0.9200	0.9975	0.0025	0.9200
5	0.8000	0.8696	0.9938	0.0062	0.8333	0.8800	0.9167	0.9963	0.0037	0.8980
6	0.8000	0.9091	0.9938	0.0062	0.8511	0.9600	0.8571	0.9987	0.0013	0.9057
7	0.8000	0.0062	0.8696	0.0062	0.9938	0.8000	0.8333	0.9938	0.0062	0.8163
8	0.7600	0.7037	0.9925	0.0075	0.7308	0.8400	0.8750	0.9950	0.0050	0.8571
9	1.0000	0.8065	1.0000	0	0.8929	1.0000	0.9615	1.0000	0	0.9804
10	0.8400	0.9545	0.9950	0.0050	0.8936	0.8800	0.8148	0.9962	0.0038	0.8462
11	0.9600	0.9600	0.9988	0.0013	0.9600	0.9600	0.9231	0.9987	0.0013	0.9412
12	0.7600	0.7600	0.9925	0.0075	0.7600	0.9200	0.9583	0.9975	0.0025	0.9388
13	0.9600	0.9600	0.9988	0.0013	0.9600	0.7600	0.9500	0.9925	0.0075	0.8444
14	0.9200	0.8846	0.9975	0.0025	0.9020	0.9200	0.9200	0.9975	0.0025	0.9200
15	0.8400	0.7778	0.9950	0.0050	0.8077	0.9600	0.8889	0.9987	0.0013	0.9231
16	0.8800	0.9565	0.9963	0.0037	0.9167	0.8000	0.8333	0.9938	0.0062	0.8163
17	0.8400	0.9130	0.9950	0.0050	0.8750	0.9200	0.8846	0.9975	0.0025	0.9020
18	0.8800	0.7857	0.9962	0.0038	0.8302	0.9200	0.8846	0.9975	0.0025	0.9020
19	0.9600	0.8276	0.9987	0.0013	0.8889	0.9600	0.9231	0.9987	0.0013	0.9412
20	0.8800	0.8462	0.9962	0.0038	0.8627	0.9200	1.0000	0.9975	0.0025	0.9583
21	0.8400	0.8077	0.9950	0.0050	0.8235	0.9600	0.8571	0.9987	0.0013	0.9057
22	0.9200	0.8214	0.9975	0.0025	0.8679	0.8800	0.8462	0.9962	0.0038	0.8627
23	0.9200	0.9200	0.9975	0.0025	0.9200	0.9200	0.9583	0.9975	0.0025	0.9388
24	0.9200	0.9583	0.9975	0.0025	0.9388	0.9600	0.8571	0.9987	0.0013	0.9057
25	0.8800	1.0000	0.9963	0.0037	0.9362	0.9600	0.9600	0.9988	0.0013	0.9600
26	0.8800	0.8800	0.9962	0.0037	0.8800	0.8400	0.9130	0.9950	0.0050	0.8750
27	0.9600	0.8000	0.9987	0.0013	0.8727	0.8400	0.9545	0.9950	0.0050	0.8936
28	0.8800	0.8462	0.9962	0.0038	0.8627	0.9600	0.9600	0.9988	0.0013	0.9600
29	0.9600	0.8889	0.9987	0.0013	0.9231	0.9200	0.9200	0.9975	0.0025	0.9200
30	0.8400	0.9545	0.9950	0.0050	0.8936	0.9200	0.9200	0.9975	0.0025	0.9200
31	0.8800	1.0000	0.9963	0.0037	0.9362	0.8400	0.8750	0.9950	0.0050	0.8571
32	0.8000	0.9091	0.9938	0.0062	0.8511	0.9200	0.7667	0.9975	0.0025	0.8364
33	0.8800	0.8462	0.9962	0.0038	0.8627	0.8400	0.9545	0.9950	0.0050	0.8936
Avg	0.8764	0.8812	0.9961	0.0039	0.8764	0.9018	0.9044	0.9969	0.0031	0.9016

4 Conclusion

In this paper, we have presented a method for Offline Malayalam Character Recognition using multiple classifier combination scheme. Chain Code Histogram and Fourier Descriptors were extracted from preprocessed character images to form feature vectors. These features were fed as input to two feedforward neural networks. The final outputs were obtained by combining the results of both these neural networks using a weighted majority scheme. The proposed system achieves a recognition accuracy of 92.84% and 96.24% for the writer independant and writer dependant schemes respectively. From the analysis of the confusion matrix obtained for the experiment, we have found that most of the misclassifications are due to confusing pair of characters. So the future work aims at reducing these errors by incorporating additional features in the post processing stage.

References

1. Plamondan, R., Srihari, S.N.: Online and offline handwriting recognition: A comprehensive survey. IEEE Trans. on PAMI 22(1), 63–84 (2000)
2. John, J., Pramod, K.V., Balakrishnan, K.: Handwritten Character Recognition of South Indian Scripts: A Review. In: National Conference on Indian Language Computing, Kochi, February 19-20, pp. 1–6 (2011)
3. Abdul Rahiman, M., Rajasree, M.S.: A Detailed Study and Analysis of OCR Research in South Indian Scripts. In: Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, pp. 31–38. IEEE (2009)
4. Chacko, A.M.M.O., Dhanya, P.M.: Handwritten Character Recognition in Malayalam Scripts -A Review. Int. Journal of Artificial Intelligence & Applications (IJAIA) 5(1), 79–89 (2014)
5. Rajasekararadhya, S.V., Ranjan, P.V.: Efficient zone based feature extraction algorithm for handwritten numeral recognition of popular south Indian scripts. Journal of Theoretical & Applied Information Technology 7(1), 1171–1180 (2009)
6. Lajish, V.L.: Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks. In: Proc. 4th Int. National Conf. on Innovations in IT, pp. 188–192 (2007)
7. Moni, B.S., Raju, G.: Modified Quadratic Classifier for Handwritten Malayalam Character Recognition using Run length Count. In: Proceedings of ICETECT, pp. 600–604. IEEE (2011)
8. Vidya, V., Indhu, T.R., Bhadran, V.K., Ravindra Kumar, R.: Malayalam Offline Handwritten Recognition Using Probabilistic Simplified Fuzzy ARTMAP. In: Abraham, A., Thampi, S.M. (eds.) Intelligent Informatics. AISC, vol. 182, pp. 273–283. Springer, Heidelberg (2013)
9. Arora, S., Bhattacharjee, D., Nasipuri, M.: Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition. In: Proceedings of: IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, India, December 8-10, pp. 1–6 (2008)
10. Freeman, H.: On the encoding of arbitrary geometric configurations. IRE Trans. on Electr. Comp. or TC(10) (2), 260–268 (1961)