Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 86 – 92

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Multiple Classifier System for Offline Malayalam Character Recognition

Anitha Mary M.O. Chacko[a,*], Dhanya P.M.[b]

[a]*Department of Computer Science & Engineering, Rajagiri School of Engineering & Technology, Kochi, 682039, India*
[b]*Department of Computer Science & Engineering, Rajagiri School of Engineering & Technology, Kochi, 682039, India*

Abstract

This paper presents a multiple classifier system for the recognition of offline handwritten Malayalam characters. The features used are the gradient and density based features. These feature sets are fed as input to two feedforward neural networks. The results of both these neural networks are combined using four different combination schemes: Max rule, Sum rule, Product rule and Borda count method. The best combination ensemble with an accuracy of 81.82% is obtained by using the Product rule combination scheme.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

*Keywords:*Character Recognition; Gradient feature; Density feature; Multiple Classifier System; and Neural Networks ;

1. Introduction

Offline recognition of handwritten text is one of the most challenging research areas due to the lack of temporal information as available in the online data and the large variations encountered in the writing style of different writers. Even though the OCR research is well advanced for foreign languages, the research on Indic scripts, especially South Indian languages is still in the infancy stage. Among the South Indian languages, the recognition of

* Corresponding author. Tel.:+91-9961573757;
*E-mail address:*anithamarychacko@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.01.061

Malayalam scripts poses an even greater challenge due to the extremely large character set and highly similar writing style of the characters.

Malayalam is one of the 4 major Dravidian languages of South India. The basic character set of Malayalam consists of 15 vowels and 36 consonants. Apart from these basic characters, the script consists of other vowel modifiers, conjunct consonants etc which together with the basic characters constitute the complete Malayalam character set consisting of 128 characters.

In[1], gradient directions along with MQDF classifier were used for the recognition of Malayalam characters. Gradient and curvature calculation along with dimensionality reduction using Principal Component Analysis was carried out in[2]. Recently, G Raju[3], proposed a Malayalam character recognition system using gradient based features and Run length count. The authors have proposed another character recognition scheme using the fusion of global and local features for the recognition of isolated Malayalam characters[4]. The authors have also applied gradient features for the recognition of Malayalam vowels in[5]. Arora et al.[6] proposed a multiple classifier system using chain code histogram and moment invariants for the recognition of Devanagari character recognition.

In this paper, we present a multiple classifier combination system for the recognition of Malayalam characters. From the preprocessed images, the gradient directions and density features are extracted which forms the input for the two feedforward neural networks. The results of these two neural networks are then combined to form the final recognition results. We have used four combination strategies : Max Rule, Product Rule, Sum Rule and Borda Count Method.

The paper is structured as follows: Section 1 gives a brief introduction about the offline character recognition domain. The architecture of the proposed system is introduced in Section 2. The experimental setup and results obtained are presented in Section 3. Finally Section 4 concludes the paper.

2. Proposed System

The proposed system consists of mainly 4 phases: Preprocessing, Feature extraction, Classification and Post Processing. The scanned image is first preprocessed to remove as much distortions as possible. In the feature extraction phase, Gradient features and Density features are extracted from 2x2 zones and 4x4 zones respectively. These features are fed as input to two feedforward neural networks and finally the results of these classifiers are combined using 4 schemes: Max rule, Sum rule, Product rule and Borda count method. The architecture of the proposed system is shown in Fig. 1.

2.1. Preprocessing

Preprocessing steps are applied in order to reduce the variations in the scanned image as much as possible. Here, the scanned image is binarized using Otsu's method of global thresholding. Segmentation is an important phase in preprocessing. The line separation and character separation is done here using horizontal and vertical projection profiles. Finally the image is normalized to 32 * 32 using bicubic interpolation.

2.2. Feature Extraction

Feature extraction is the most important phase in character recognition which determines the success rate of the underlying OCR system. Here, two feature sets are extracted in this phase: Gradient features and density features.

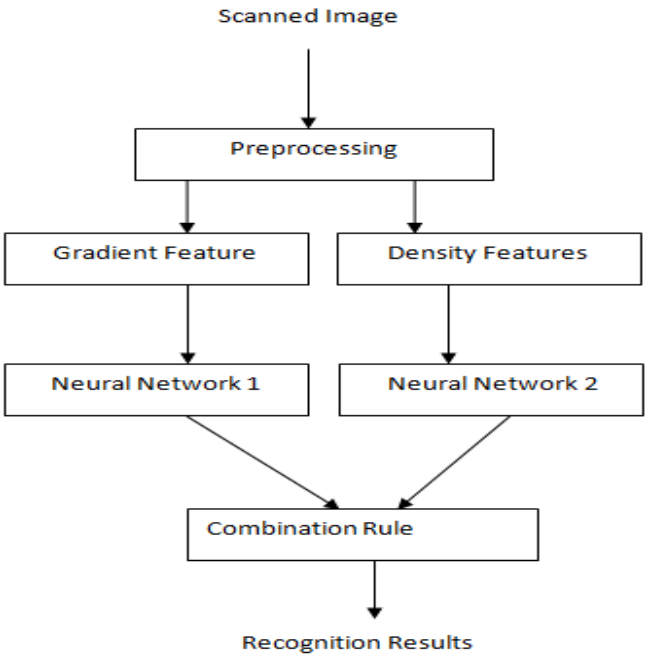

Fig. 1. Proposed System Architecture

2.2.1 Gradient Features

Gradient features at each pixel position of an image points in the direction of greatest rate of change of intensity and thus represents the local characteristics of the image. The gradient features have been successfully applied for Malayalam character recognition[1-5]. Here, the preprocessed images are divided into 2 x 2 zones and then gradient directions are computed for each of the zones using 3 x 3 sobel operators. Sobel operator uses two templates for the calculation of gradient in the horizontal and vertical directions as in Fig. 2.

a

1	2	1
0	0	0
-1	-2	-1

b

-1	0	1
-2	0	2
-1	0	1

Fig. 2. Sobel Operator (a) Horizontal Template; (b) Vertical Template.

Based on these templates, the gradient components at pixel position (i,j) are calculated as:

$$g_h(i,j) = f(i-1,j-1) + 2f(i-1,j) + f(i-1,j+1) - f(i+1,j-1) - 2f(i+1,j) - f(i+1,j+1)$$

$$g_v(i,j) = f(i-1,j+1) + 2f(i,j+1) + f(i+1,j+1) - f(i-1,j-1) - 2f(i,j-1) - f(i+1,j-1)$$

The gradient direction at location (i,j) are then calculated from these components as :

$$\theta = \arctan \frac{g_V(i,j)}{g_h(i,j)}$$

The gradient directions at each pixel are mapped into 8 directional codes with an equal angle span of 45 degrees between each of them. The sum of each direction codes from each of the 4 blocks form 32 features (Feature set I).

2.2.2 Density Features

The pixel density feature is a simple yet effective feature for the recognition of Malayalam characters. The pixel density feature is computed as the ratio of number of foreground pixels in each zone to the total number of pixels in that zone. For computing density features, the preprocessed character images are divided into 4 x 4 zones. Then for each of the 16 zones, the density features are calculated as:

$$D(i) = \frac{\text{Number of pixels in zone i}}{\text{Total number of pixels in zone i}}$$

Thus the 16 density features form Feature Set II.

2.3. Classification

Classification is the final phase in character recognition that assigns a unique label to character images based on the features extracted. Here we have used two feedforward neural networks trained with the 32 gradient features and 16 density features respectively. Each of the neural networks are trained with resilient backpropagation algorithm. The final recognition results are obtained by combining the results of these individual networks.

2.3.1 Classifier Combination Methods

Combining the results of individual classifiers leads to better recognition results as the set of patterns misclassified by different classifiers would not necessarily overlap[7]. Here we have used four combination strategies for combining the results of our individual classifiers: Max rule, Sum rule, Product Rule and Borda Count method.

The max rule selects the class with the maximum confidence value among both classifier as the final output. The sum rule sums up the confidence values for each class and selects the one with the highest sum as the recognition result whereas the product rule multiplies the confidence values for each class and selects the one with the highest value. In the borda count method, the classes are sorted in descending order of their borda count values and the class with the highest borda count value is selected as the output class. The Borda count for a class c is the sum of number of number of classes ranked below the class by each classifier.

3. Experimental Results

The proposed system was implemented using Matlab R2013a. The experiments were conducted on 33 selected characters of Malayalam – 8 isolated vowels and 25 consonants. The database used for the experiment consists of 825 samples of the above characters collected from 25 different people belonging to different age groups and professions. Samples of the characters used for this experiment are given in Fig. 3.

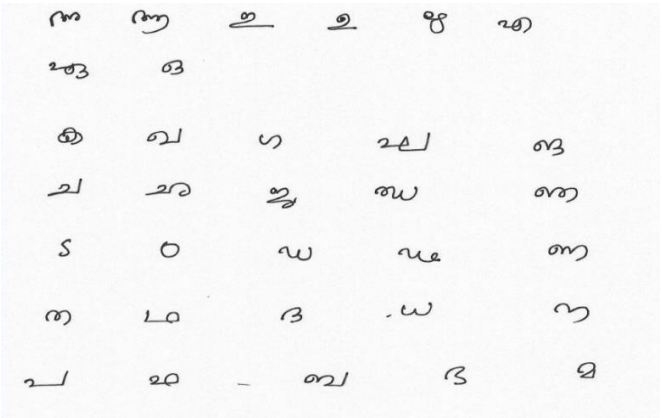

Fig. 3. Character Samples

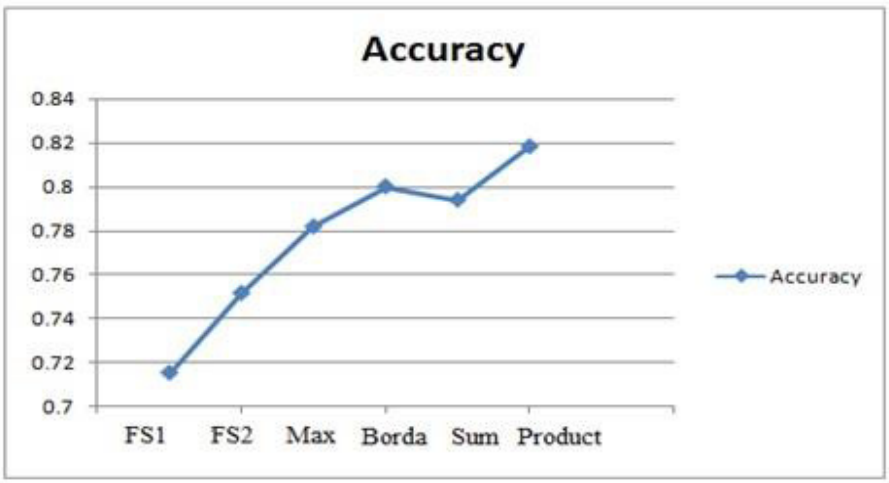

Fig. 4. Performance of different classifier systems

From the database, 80% of the samples were used for training and the rest were used for testing. The recognition accuracy obtained from the gradient feature based classifier and density based classifier are 75.15% and 71.52% respectively. The combined classifier system using the Max rule, Borda Count Rule, Sum Rule and Product Rule achieves recognition accuracies of 78.18%, 79.39%, 80% and 81.82% respectively. These results are summarized in Fig. 4.

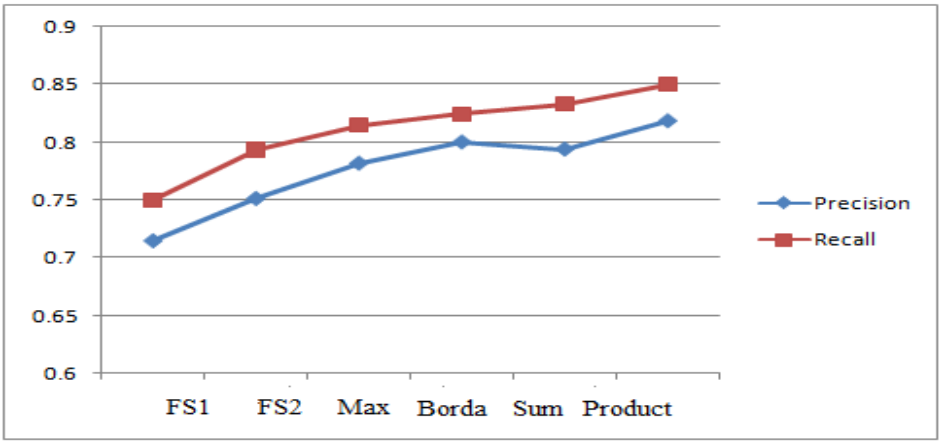

Fig. 5.Precision and Recall

Based on the confusion matrix of the experiment, we have computed several useful measures such as average accuracy, precision, recall, Fscore, Specificity and FP Rate. These results are summarized in Table 1.The precision and recall graph for each of the classifier system are summarized in Fig. 5.

Table 1.Classification Results

Classifier System	Accuracy(%)	Precision	Recall	Fscore	FP Rate	Specificity
Density	71.52	0.7152	0.7497	0.7077	0.0088	0.9912
Gradient	75.15	0.7515	0.7934	0.7436	0.0077	0.9923
Max	78.18	0.7818	0.8146	0.7730	0.0068	0.9932
Borda	79.39	0.7939	0.8332	0.7920	0.0064	0.9936
Sum	80	0.8	0.8242	0.7882	0.0062	0.9938
Product	81.82	0.8182	0.8499	0.8153	0.0056	0.9944

4. Conclusion

In this paper, we have presented a multiple classifier system for the recognition of offline Malayalam characters. Gradient features and density features were extracted from preprocessed character images to form feature vectors which were fed as input to two feedforward neural networks. The final results were obtained by combining these two neural networks using 4 combination strategies: Max rule, Borda count method, Sum rule and Product rule. The

proposed system achieves a recognition accuracy of 81.82% using the Product rule combination scheme. All these schemes have given equal importance to both the individual classifiers. So, our future work aims at finding other effective combination schemes that assigns weight to these individual classifiers depending on the individual classification results.

References

1. Bindu S Moni and G. Raju, Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition, *IJCA Special Issue on Computational Science- New Dimensions & Perspectives* (1), 2011, p. 30-34.
2. Jomy John, Kannan Balakrishnan, Pramod K. V, *A* System for Offline Recognition of Handwritten Characters in Malayalam Script, *I.J. Image, Graphics and Signal Processing*, 2013. 4.p.53-59.
3. G Raju, Bindu S Moni, Madhu S. Nair, *A Novel* Handwritten Character Recognition System Using Gradient Based Features and Run Length Count*, Sadhana Indian Academy of Sciences*, Springer India, 2014, p. 1-23.
4. Anitha Mary M.O. Chacko, Dhanya P.M., *Offline* Malayalam Character Recognition Using Global and Local Features*,* Proceedings of the Second International Conference on Emerging Research in Computing, Information, Communication and Applications, *ERCICA*, 2014, Elsevier Publications, p. 806-812.
5. Anitha Mary M.O. Chacko, Dhanya P.M, Gradient and Density Based Features for Offline Malayalam Character Recognition*,* National Conference on Computer Competence, *NCCC'14*, 2014, p 1-4.
6. Arora Sandhya, Bhattacharjee Debotosh, Nasipuri Mita, Basu Dipak Kumar and Kundu Mahantapas, Combining multiple feature extraction techniques for handwritten Devnagari character recognition, Third international Conference on Industrial and Information Systems *ICIIS*, 2008, p. 1-6.
7. Joseph Kittler, Mohamed Hatef, Robert P.W. Duin and Jiri Matas, On Combining Classifiers, IEEE *Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, No 3. March 1998, p. 226-239.