Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 25 (2016) 224 – 231

Procedia Technology

Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science and Technology (RAEREST 2016)

An Approach towards Malayalam Handwriting Recognition Using Dissimilar Classifiers

Meenu Alex[a*], Smija Das[b]

Department of Computer Science & Engineering, St. Joseph's College of Engineering & Technology, Palai, Kerala,India-686579

Abstract

The handwriting recognition in Malayalam is a challenging as well as emerging area of pattern recognition. It is a tedious process mainly due to its enormous character set. Here we propose a novel method for handwriting recognition by using two dissimilar classifiers. It can also be called as an ensemble method in which multiple classifiers are combined to solve a particular problem and thereby improve the performance of the system. The experiment is conducted in 2 phases. In the first phase, 33 isolated characters in Malayalam were used. In the second phase, Malayalam sentences were used. From the preprocessed image, we were extracted two features: SURF feature and Curvature feature. These features were fed as input to a neural network and an SVM classifier. Finally, the result of both the classifiers was combined to get the final results. The system showed an accuracy of 89.2% in the first phase. An accuracy of 81.1% was exhibited in the second phase.

Keywords: Handwriting Recognition; Segmentation; SURF;Curvature feature; SVM; Neural Network; Ensemble learning

1. Introduction

Now is the age of digitalization. Rather than storing the data on papers, it can be safely stored and easily accessed, then digitalized. Therefore Handwriting Recognition is of prime importance. Handwriting Recognition is an emerging as well as challenging area in the fields of pattern recognition and computer vision. Various Indian offices including post offices, banks etc handle lot of handwritten documents and thus the handwriting recognition (HCR) has attained a hotspot area of research. Other applications include automatic number plate recognition, CTS scanning, preservation of degraded documents etc. The aim of a handwriting recognition system is to convert human readable characters which are present in a photographed or digitized sheet of paper and convert it into a machine editable form. A handwriting recognition system is of two types: online or offline. In an online character recognition system, the data is captured at the same time when the user writes on a digitizer. It actually does the real time conversion of characters

doi:10.1016/j.protcy.2016.08.101

to their corresponding Unicode values. In an offline system, a scanner captures the data after the writing process is over.

The recognition of printed text is a comparatively simple process. The variation in between a character printed at different parts of the world is negligible. But the scenario is extremely different in the case of handwritten texts. A character being sketched in hand differs individually depends upon a number of factors. The emotion, mood, age, health condition etc of the writer along with the variation provided by the pen used. All makes numerous variations in a character being written by different individuals at different parts of the world. The complexity of our problem thus lies in the complexity of our language added up to the huge variations in penning down the characters that vary individually.

An important feature of Malayalam language is its enormous character set. Thus the identification of characters may be posing another challenge in the form of similarity between the characters. Adding up to the scene is the similarity in writing styles of different people. Character recognition has already been successful in foreign languages like English, Japanese, Chinese, Arabic etc. The universal language English presents itself simple with only 26 characters – not forgetting the divisions based on the case of letters. Several researchers have come out with notable accuracy in the handwritten recognition of English by virtue of this.

The recognition of scripts is a tedious process for South Indian languages like Malayalam, Kannada, Tamil, Telugu etc. This is mainly due to the large character set, presence of compound characters and so on. Malayalam, being the second most difficult language in the globe, has a distinguished character set varying largely in between them. So to reach 100% accuracy is very difficult job. The scope of this work lies in converting important handwritten Malayalam documents into well kept digital data.

2. Related Work

A lot of works are reported in foreign languages [2] in the domain of Handwritten Character Recognition. Among Indian languages like Devanagari, Tamil, Oriya and Bangla many works were occurred [3,4]. But in case of South Indian languages especially Malayalam, only few works were reported. A recognition system that can identify the complete character set of Malayalam is not developed till now. The main difficulty in Malayalam character recognition system is the lack of availability of a benchmarking database for comparison.

The first work in Malayalam Character Recognition was reported in 2007 by Lajish V.L [5]. It uses fuzzy-zoning method and normalized vector distance measures for the recognition of 44 Malayalam characters. Renju John et al [6] came forward with the concept of 1D Wavelet Transform of Projection Profiles for Isolated Handwritten Malayalam Character Recognition. Handwritten character recognition by applying Daubechie wavelet coefficients was proposed in [7]. M. Abdul Rahiman et al [8] proposed an HLH intensity pattern based method for recognition of Malayalam characters. A recognizing system for Malayalam characters using discrete features was introduced by Binu P. Chacko and Babu Anto in [9]. Jomy John et al [10] proposed a chain code histogram based method for recognizing vowels of Malayalam. Bindu S. Moni et al [11] invented a handwriting recognition system based on run length count (RLC). Vidya V. et al [12] proposed a method for handwritten character recognition based on Probabilistic Simplified Fuzzy ARTMAP (PSFAM). Features like Zernike moment features, cross feature, distance feature and fuzzy depth are extracted from the character. This method gained an accuracy of 79.48% for 142 Malayalam characters. In 2014, Shanjana C et al [13] proposed a method for Malayalam character segmentation. In this work, segmentation of characters is performed by combining Vertical projection profile method along with connected component analysis method.

The above mentioned HCR systems propose a variety of feature extraction methods but a good result has not yet been achieved till now. All the above mentioned works are based on a single classifier system. Even though combination of multiple classifiers (similar classifiers) has been experimented in Malayalam, but a combination of dissimilar classifiers has not been tested till now. This motivated us to examine the outcome of using dissimilar classifier combination in Malayalam.

3. Characteristics of Malayalam Scripts

Malayalam is one among the regional languages in India which owes its origin to Sanskrit. Designated as a classical language in 2013, it has the reputation of being the second most difficult language to be proficient with. It is mainly used in the state of Kerala, Union territory of Lakshadweep and Mahe. Malayalam script is derived from Grantha script.

The letters of Malayalam script consists of curves and loops. The vastness of character set is yet another distinguishing factor of Malayalam. Malayalam characters can be basically categorized as vowels and consonants. It also contains 9 rarely used numerals. Another division of Malayalam Character comprises of the conjunct consonants, compound characters and consonant diacritics[21].

അആഇ ഈഉ ഊഋ എ

ഏ ഐ ഓ ഔ□□ അ

Figure 1: Malayalam Vowel Set

ക ഖ ഗ ഘങ

ച ഛജ ഝഞ

ട ഠ ഡഢണ

ത ഥ ദ ധ ന

പ ഫ ബഭ മ

യ ര ല വ ശ

ഷ സ ഹ ള ഴ റ

Figure 2: Malayalam Consonant Set

4. Proposed Method

The proposed system consists mainly of the stages given below.

- Image Acquisition
- Preprocessing
- Segmentation
- Feature Extraction
- Classification
- Post Processing

The layout of the system is given in figure 3. Initially the samples were collected from people of different age group, sex and profession. Then the image is subjected to preprocessing to remove distortions present in the image. After that line segmentation and character segmentation were performed to isolate the individual characters. Then SURF feature, curvature feature and diagonal feature were extracted from the image. Finally, two dissimilar classifiers namely SVM and neural networks were used for classifying the features obtained. SVM was trained with SURF and curvature feature. Neural network was trained with diagonal feature. Then results from both the classifiers were then combined to get the final recognition result. The obtained result was then converted to Unicode values in the post processing stage.

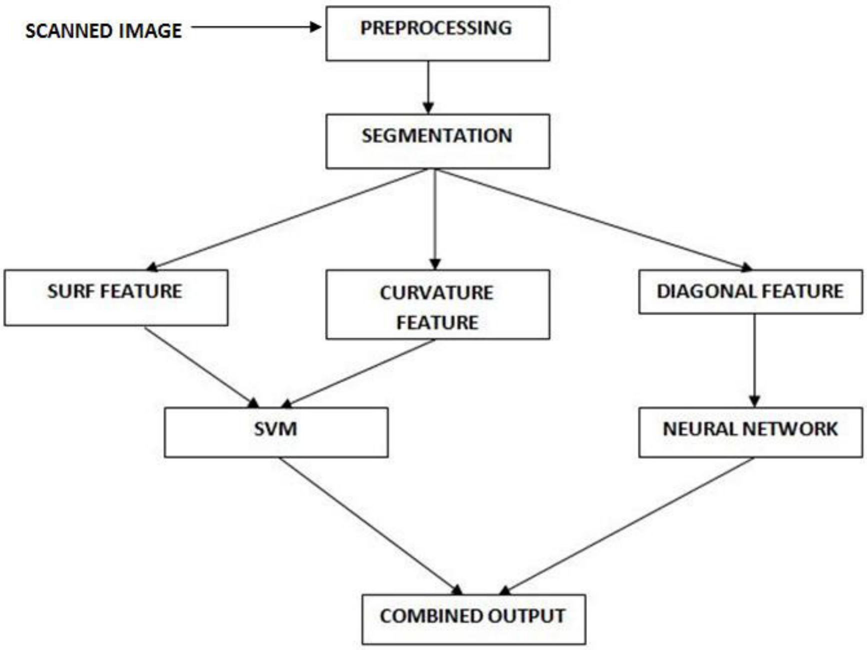


Figure 3: Proposed Method

4.1Image Acquisition

The images for the experiment are captured by using a scanner at 300 dpi resolution. The images can be in any format like JPEG, BMP, PNG etc. These images are given as input to the system for further steps. A sample scanned image is shown in figure 4.

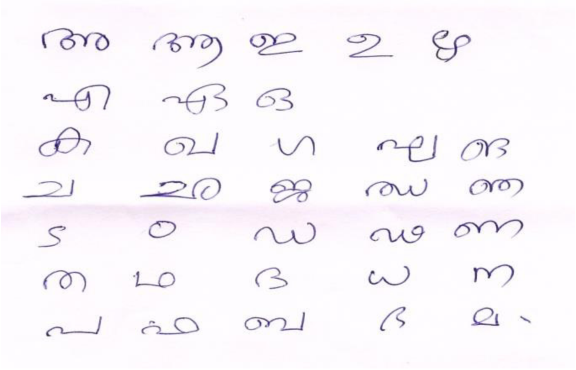


Figure 4. Character sample

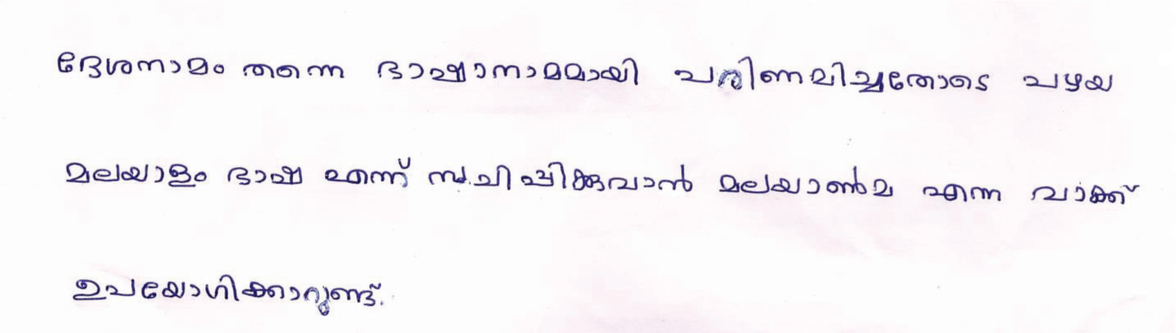

ദേശനാമം തന്നെ ഭാഷാനാമമായി പരിണമിച്ചതോടെ പഴയ
മലയാളം ഭാഷ എന്ന് സൂചിപ്പിക്കുവാൻ മലയാണ്മ എന്ന വാക്ക്
ഉപയോഗിക്കാറുണ്ട്.

Figure 5 : Sentence sample

4.2.Preprocessing

The aim of preprocessing is to remove as much as distortions as possible from the scanned image. Degraded documents or poor quality of scanner are responsible for these distortions. At first, the scanned image is converted to grayscale if it is in RGB or any other format. After that the grayscale image is converted to binary using Otsu's method of global thresholding. The advantage of using Otsu method is that no prior knowledge about the image is required for the conversion. In order to remove the salt and pepper noise present in the image, a 3x3 median filter is used. Thinning is performed using Hilditch algorithm to extract only the relevant features from the image.

4.3.Segmentation

The purpose of segmentation is to isolate individual characters from the handwritten image. The task of segmentation from handwritten text is complex due to the existence of broken characters, touching characters and overlapped characters. Line segmentation from text is performed by using Horizontal projection profile. Character segmentation is done by Vertical Projection Profile Method. An example of touching character is shown in figure 6.

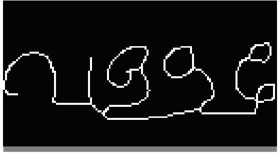


Fig 6: Touching Character

4.4.Feature Extraction

The idea behind performing feature extraction is to extract the salient characteristics of the image. The success of a character recognition system relies on effective feature – classifier combination. Here we were extracted two features from the character. SURF and curvature are the extracted features. These extracted features are given as the input to two classifiers.

4.4.1. SURF- It is called Speeded Up Robust Features. It is the speeded up version of SIFT. The main benefit of using SURF in the feature extraction stage is its capability to distinguish between points of interest in the image. SURF Algorithm involves 3 parts: Interest Point Detection, Local Neighborhood Description and Matching.

4.4.2 Curvature Feature- Majority of the Malayalam letters contains curves and loops. Since most of them are round and curved letters, curvature feature is suitable to identify and extract the features from them. Curvature is a feature that provides clearer structural sketch of the image.

4.4.3 Diagonal Feature- In diagonal feature extraction method, the input image is divided into zones of predefined sizes. Then features are computed for each of these zones. Zoning provides the local characteristics of an image. Here the image of size 90x60 is divided into 54 zones in which each zone is of size 10x10 pixels. Each zone provides 19 subfeatures which are averaged to form a single feature vector and stored on its corresponding zones. This process is repeated for each zone and finally we get 54 features for each character.

4.5.Classification

The decision making part as well as the final stage of a character recognition system is classification. In this stage, unique labels are assigned to each character image based on the extracted features. Here we have used a pair of dissimilar classifiers. SVM and Neural Network are the classifiers used here.

4.5.1 SVM (Support Vector Machine) - It is a supervised learning model which performs linear classification. The advantage of using SVM is that it learns fast. But on the other side, it predicts slowly. SVM is used when your data has extremely 2 classes. SVM classifies data by finding the best hyperplane that separates all datapoints of one class from another.

4.5.2 *Neural Network* - It is a network that learns from observed data. Neural Network learns slow but has fast prediction capacity. This can perform tasks which cannot be performed by a linear program. Neural Network for Malayalam character recognition consists of a set of input neurons that are activated by the pixels of an input image. The activation is then passed onto other set of neurons. The process is repeated until an output neuron is activated. This determines which character was read.

4.6. Post Processing

In the post processing stage, the characters are mapped to their corresponding Unicode values. In Malayalam, certain characters like compound characters do not have Unicode values. So we need to manually find out their Unicode values. For eg: □□ + ത= □□□

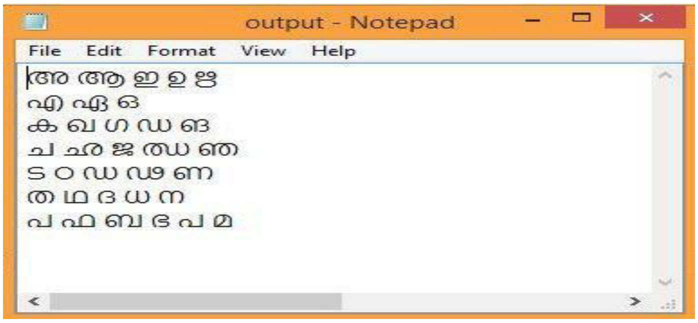
5. Combining Dissimilar Classifiers

Neural Network and SVM are very good classifiers for classification purpose. We get competitive results from both classifiers. So better results can be obtained by aggregating the results. Furthermore, dissimilar classifier combination overcomes the limitations of individual classifiers. By using the method proposed above, we get 2 outputs. We are combined the output to get the final result.

Both classifiers have their own benefits and limitations. Combination of dissimilar classifiers will yield more productive results. A proper combination of dissimilar classifiers will give more accurate results because each classifier provides complementary information about the pattern to be classified. When dissimilar classifiers are used, they will complement each other on the inputs disagreeable to either one of them.

6. Results and Discussions

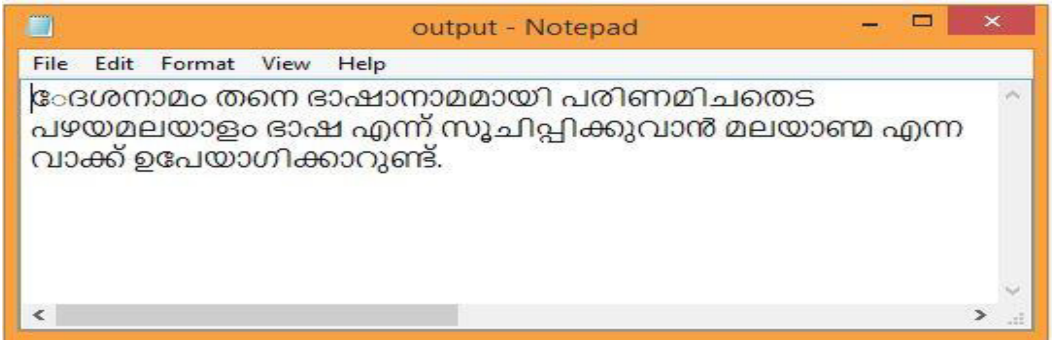
The experiment was conducted on 2 phases. We have used 300 dpi digitized images for the experiment. In the first phase, we were asked the people to write 33 Malayalam isolated characters. We have collected data from 25 people who were of different gender, age group and profession. No restriction is imposed on the writers for the writing purpose. After that we have performed the stages mentioned above. Segmentation is not much important in this phase. We are only performing character segmentation. This method showed an accuracy of 89.2% for 33 Malayalam character classes.



```
അ ആ ഇ ഉ ഋ
എ ഏ ഒ
ക ഖ ഗ ഘ ങ
ച ഛ ജ ഝ ഞ
ട ഠ ഡ ഢ ണ
ത ഥ ദ ധ ന
പ ഫ ബ ഭ പ മ
```

Fig 7: Recognition of Alphabets

In the second phase, people were asked to write Malayalam sentences. Segmentation has major role here due to the proximity of lines, touching and overlapped characters. This method recognize independent and isolated characters, characters containing *chandrakkala*, *valli* etc. After segmentation, the characters were subjected to SURF, Curvature and diagonal feature extraction. The output of feature extraction was fed as input to the two classifiers. This phase attained 81.1% accuracy, which is less then the first phase.



```
േദശനാമം തനെ ഭാഷാനാമമായി പരിണമിചതടെ
പഴയമലയാളം ഭാഷ എന്ന് സൂചിപ്പിക്കുവാൻ മലയാണ്മ എന്ന
വാക്ക് ഉപേയാഗിക്കാറുണ്ട്.
```

Fig 8: Recognition of Sentences

7. Conclusion

In this paper, we propose a method for the recognition of Malayalam handwritten characters using a combination of dissimilar classifiers. The work was conducted in 2 phases. Severe accuracy reduction can be seen from first phase to second phase. This is mainly due to the presence of touching characters. In order to identify individual characters, we extract SURF, curvature and diagonal feature and for classifying Neural Network and SVM are used. As an extension to this work, we can think of a system that can recognize characters from old degraded documents like legal documents and convert it into machine editable form in order to preserve them.

Acknowledgements

I take this opportunity to thank Ms. V. Vidya, Senior Engineer at CDAC, Thiruvananthapuram who has supported me whole heartedly for the successful completion of this work.

References

[1] R. R. Plamondan, S.N. Srihari, “Online and offline handwriting recognition: A comprehensive survey”,IEEE Trans. On PAMI, Vol 22(1) pp 63-84,2000.

[2] Jannoud, I.A.: “Automatic Arabic Hand Written Text Recognition System”. American Journal of Applied Sciences (2007)

[3] Ved Prakash Agnihotri “Offline Handwritten Devanagari Script Recognition”, I.J. Information Technology and Computer Science, 2012, 8, 37-42

[4] Pal, U., et al.: “Handwritten Bangla Compound Character Recognition Using Gradient Feature”. In: 10th International Conference on Information Technology (2007)

[5] 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, pp 188-192

[6] R. 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 13-15K.

[7] 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.

[8] M. A. Rahiman et. al., “Isolated handwritten Malayalam character recognition using HLH intensity patterns”,Second International Conference on Machine Learning and Computing,2010

[9] Binu P. Chacko, Babu Anto P, “Discrete Curve Evolution Based Skeleton Pruning for Character Recognition”, Seventh International Conference on Advances in Pattern Recognition, 2009.

[10] Jomy John, Pramod K. V, Kannan Balakrishnan ”Offline Handwritten Malayalam Character Recognition Based on Chain Code Histogram”, Proceedings Of ICETECT 2011.

[11] Bindu S Moni, G Raju, “Modified Quadratic Classifier for Handwritten Malayalam Character Recognition using Run length Count”, In International Conference IEEE, 2011.

[12] Vidya V, Indhu T R, Bhadran V K,R Ravindra Kumar, “Malayalam Offline Handwritten Recognition using Probabilistic Simplified Fuzzy ARTMAP”, Advances in Intelligent Systems and Computing Volume 182, 2013, pp 273-283.

[13] Shanjana C, Ajay James, “Character Segmentation in Malayalam Handwritten Documents”, IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), August 2014

[14] B. Anuradha and B. Koteswarra; “An efficient Binarization technique for old documents”, Proc.of International conference on Systemics,Cybernetics and Inforrmatics,Hyderabad, pp 771-775,2006

[15] Bindu S Moni, G Raju, “Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition”, IJCA Special Issue on Computational Science - New Dimensions Perspectives NCCSE, 2011

[16] Jomy John, Pramod K. V., Kannan Balakrishnan, ”Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier”, In International Conference oncommunication Technology and System Design, ELSEVIER 2011.

[17] 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.

[18] Abdul Rahiman M, M. S. Rajasree, Masha N, Rema M ,Meenakshi R, Manoj Kumar G, ”Recognition of Handwritten Malayalam Characters using Vertical Horizontal Line Positional Analyzer Algorithm”, IEEE International Conference 2011.

[19] Arica, N., Yarman-Vural, F.T.: “An Overview of Character Recognition Focused on Off-Line Handwriting”. IEEE Transactions on System, Man and Cybernetics –Part C: Applications and Reviews (2001)

[20] Pal, U., et al.: “A System for Off-line Oriya Handwritten Character Recognition using Curvature Feature”. In: 10th International Conference on Information Technology (2007)

[21] Meenu Alex, Smija Das: “A Study on offline character recognition in Malayalam scripts”. Proc. Of International Conference on Emerging Trends in Technology and Applied Sciences.p. 47,April-May 2015.

[22] Meenu Alex, Smija Das:” Offline Malayalam Character Recognition using Genetic Algorithm”.International Journal of Advanced Technology in Engineering & Science, Vol No 3, Special Issue No 1, September 2015

[23] Bay H, Tuytelaars T, van Gool L J. SURF: “Speeded Up Robust Features”[C]//European Conference on Computer Vision, 2006, I:404-417.