Intelligent Handwritten Character Recognition For Malayalam Scripts Using Deep Learning Approach

Bineesh Jose¹, K. P. Pushpalatha²

¹ Research Scholar, School of Computer Sciences, M. G. University, Kottayam, Kerala, India bineeshjose@gmail.com
² Associate Professor & HoD, School of Computer Sciences, M. G. University, Kottayam, Kerala, India pushpalathakp@mgu.ac.in

Abstract. Machine Learning, especially Deep Learning has been incorporated into Pattern Recognition and Image Processing for the Handwritten Character Recognition (HCR) research which is always a hot area of research. Foreign scripts are enriched with Handwritten Character Recognition (HCR) studies. But very less research can be seen for HCR in Indian scenario, especially in Malayalam script. This paper provides an overview of different feature extraction and classification techniques used for Handwritten Character Recognition of Malayalam Scripts using Deep Learning Approaches. The accuracy measures of classification of the different algorithms have been given so that the researchers can select a particular algorithm with high performance.

1. General Character Recognition System:- An Architecture View
The recognition of characters contains preprocessing, segmentation, feature extraction, and classification as shown in Fig.1.

![Figure 1. Architecture of a character recognition system](image)

1.1. Preprocessing
Preprocessing contains the following sequence of operations[11]
- Noise Removal
• Thresholding
• Skeletonization

1.2. Segmentation
It is having a major role in handwriting character recognition problem which requires line, character and word segmentation. White space and pitch, projection analysis and connected component labeling are the certain techniques used for character segmentation[11].

1.3. Feature Extraction
Feature Extraction is one of the important step towards handwriting character recognition. Salient Features are extracted from segmented images and it is captured in feature vectors. Lot of feature extraction techniques are available for character recognition system[11].

1.4. Classification
This is the final step towards the handwriting character recognition. Wide variety of classifiers are available in pattern recognition and image processing area. Any of the classification model can be used to map features to corresponding character. Supervised and Unsupervised methods are used here. Slicing of single data set into training set and test set is an important concept in Classification [6].

2. Malayalam Handwritten Character Recognition:- A Literature Review
Malayalam is one of the popular languages of South India which is mainly used by Kerala and Union territories of Lakshadweep and Mahe as official language. But very less research can be seen for HCR in Indian scenario, especially in Malayalam script. It is very difficult to have a complete handwritten text recognition system in Malayalam because of its huge character set.

2.1. HCR using Stacked Long Short Term Memory (LSTM)
2.1.1. Proposed work Jino, et al. in [1] introduced a stacked LSTM model for isolated Malayalam handwritten character recognition. 18000 samples of ninety Malayalam characters are included in the dataset. LSTM model consists of two layers and prediction is done by a final output layer. More than 90% of accuracy was achieved. Improvement can be done on dataset by adding more samples to the dataset. In this model a version of RNN which is capable of handling machine learning application called Long Short Term Memory (LSTM) was used.

LSTM Cell Architecture  LSTM Cell contains three gates namely input gate, forget gate and output gate [1] as in Fig.2 which is able to solve the vanishing gradient problem of RNN.

Stacked LSTM Architecture  The architecture Contains 3 layers, namely LSTM Layer1, Layer2 and a third Output Layer for prediction. Softmax function is used as activation function at Output prediction Layer.

2.1.2. Result  Experiment results obtained by using Stacked LSTM architecture was really good. Better results can be achieved by using more samples per class. By fine tuning the architecture parameters, this approach can lead to state of the art result. Word recognition also can be implemented by using this model.
2.2. Online Handwritten Bangla Character Recognition using CNN

2.2.1. Proposed Work Shibaprasad in [2], proposed a online handwritten isolated Bangla characters recognition system using Convolutional Neural Network (CNN) architecture. The author was tried to analyse CNN architecture with various parameters like kernel variations, pooling strategies and activation functions. He has used a dataset with 10000-character samples. The total dataset was sliced into test set and train set with a percentage of 30 and 70 respectively. An accuracy of 99.4% has been achieved on test dataset using this technique. As shown in Fig.3, the proposed system consists of 2 convolutional layers, 2 pooling layers and one fully connected classification layer.

![Figure 3. Architecture of CNN used in the present work](image)

2.2.2. Result In this paper, they have analyzed different parameters like kernel size, pooling scheme, and activation function for online handwritten Bangla character recognition

2.3. HCR using wavelet energy and extreme learning machine

Binu, et.al. in [3] employed wavelet energy feature (WEF) for feature extraction and extreme learning machine (ELM) for classification of handwritten Malayalam characters. The proposed model makes use of a new parameter called wavelet energy (WE). Characters can be represented in terms of WE distribution which returns WEF. Character images can be recognised by using WE of different decomposition levels as shown in Fig.4. These features will act as input to classifiers of handwritten character recognizer.

![Figure 2. Block Diagram of LSTM Cell](image)
Feature extraction using wavelets

Different parameters such as pixels, shape data, or mathematical properties act as features to the recognizer. In this paper WT act as a feature extractor. Pattern recognition and image processing uses wavelet transformations for the past two decades [3].

![Figure 4. 2D discrete wavelet transform](image)

Extreme learning machine

Feed-forward neural networks is consuming long time for learning which is the major disadvantage of such methods in their applications during the last 20 years. Iterative changing of parameters will lead to this issue while training such models [3].

2.3.1. Result

High learning speed and low training error are the major features of ELM with different activation functions. Wavelet energy based feature vectors are used for pattern classifications. Multiresolution analysis can be done by using WEF which is defined in terms of wavelet. More experiments can be conducted based on different features with good classifiers for better result on Malayalam character recognition.

2.4. HCR using Gradient Based Features and Run Length Count(GBF-RLC)

2.4.1. Proposed Work

Raju, et.al. in [4] proposed a model with gradient-based features and run length count (GBF-RLC) as feature extractor for handwritten character recognition system. Simplified Quadratic Classifier (SQDF) and Multi-Layer Perceptron (MLP) were used as classifiers here.

Feature extraction

The feature extraction methods are applied to each character image in the database. Character codes can be obtained from the binarized and thinned images. Binarized, resized (72x72) and thinned images are used for centroid calculation. Different steps in the feature extraction is given in Fig. 5.

Feature extraction Algorithm

It is used for extracting suitable features by applying each image in the database.
2.4.2. Result The proposed model use a combination of gradient based features as well as run length count (GBF-RLC) for hand written character recognition system. For classification, SQDF and MLP are used. In order to achieve higher recognition rate, we can use combination of features as input. The experimental result showed by SQDF and MLP classifiers are 99.66\% and 99.78\% respectively. Performance of the different methods are shown in Fig. 6.

2.5. HCR Based on Chain Code Histogram

2.5.1. Proposed Work Jomy, et.al. in [5] proposed a model for handwritten character recognition with CCH as feature extractor and Neural Network as classifier. The biggest challenge in the Malayalam Character recognition system is the largest number of characters.
The problem is more complex when we are considering different writing patterns of different people.

**Chain code calculation of Handwritten characters** Significant features are extracted from boundary points of binary segmented character image. Fig. 7 Shows the four directional and eight directional coding of segmented image. Chain code representation of the contour can be used for CCH calculation. Reconstruction of Character image using boundary points is shown in Fig. 8.

![Figure 7. Directions of four connected and eight connected chain code](image)

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

![Figure 8. Reconstruction of Character image using CCH](image)

**Figure 8. Reconstruction of Character image using CCH**

**Classification** For classification they used a Neural Network Model which is having one input layer, one output layer and many hidden layers.

2.5.2. Result The dataset is sliced into three sets as train, test and validation with percentage 70, 15 and 15 respectively. Feature set IV with 8dir CCH, 8dir NCCH and Centroid got highest average accuracy.

2.6. HWR Using a Deep Architecture

2.6.1. Proposed Approach Jino, et.al. in[7] came up with a offline model for Malayalam handwritten word recognition. He has developed a bench mark database for offline handwritten word samples. A deep convolutional neural network (CNN) architecture as in Fig. 9 was introduced for feature extraction. As usual Support vector machine (SVM) was used for classification in the above work.

The following sequence of operations are essential for the proposed word recognition model [7].

- Preprocessing
- Convolutional Neural Network
- Training of the CNN
- Support Vector Machine

![Figure 9. Proposed Architecture of Convolution Neural Network(CNN)](image)

### 2.6.2. Result
In this work the author is mainly concentrating on offline handwritten Malayalam word recognition. He is also trying to extent the same on some existing databases and got good result on this databases.

### 2.7. CNN based common approach to HCR

#### 2.7.1. Proposed Work
Durjoy, et.al. in [8] proposed a 5-layer model which is similar to LeNet-5 with two convolution layer, two subsampling layer and one fully connected layer as feature extractor. Support Vector Machine (SVM) was used as the corresponding classifier in this paper.

Accuracy can be increased by applying some preprocessing methods on the input images for handwriting character recognition. The model do not require any preprocessing to remove noises present in the input samples. The CNN architecture itself will take care of several sources of variation among the data set.

**5 Layer Convolutional Neural Network** The architecture proposed here consists of two convolutional layer, two subsampling layer and one fully connected layer [8] as shown in Fig.10.

![Figure 10. 5 Layer Convolutional Neural Network Architecture](image)

#### 2.7.2. Result
Six different character databases are used to test handwritten character recognition system in this proposal. The Proposal has shown a good recognition accuracy on train and test dataset. Any character recognition problem can reuse the feature extraction strategy that they used here. In the given proposal, they used the popular SVM classifier. By the aid of this multilayer perceptron, a moderate recognition rate is attained.
2.8. HDR Using an Improved Deep CNN Architecture

2.8.1. Proposed Work Chandrika, et.al. in [9] proposed a model for Bangla handwritten digits recognition by the aid of Deep Convolutional Neural Network (D-CNN). The D-CNN model consists of total seven layers in which three are convolution layers, another three are average pool layers and one fully connected layer for classification of Bangla handwritten digits.

2.8.2. Result The model was given 96.80% accuracy on a LeNet-5 like five-layered architecture. The total dataset for training was divided into 80 batches of 50 images each.

2.9. HCR using Neural Networks: Feature based Classification

2.9.1. Proposed Work Shreya, et.al. in [10] has done a detailed review of Handwritten Character Recognition using Neural Network by back propagation algorithm. Character recognition techniques recognize the characters written on a paper documents and convert it in digital form.

Neural Network Classification and Recognition tasks in this model are carried out by an Artificial Neural Network. In a feed forward neural network, nodes are organized in layers; each over other in a stack structure. The neural network consists of one input layer, one or more hidden layers and an output layer. Here Back propagation algorithm is used for training of multi-layer feed-forward neural network. Improved Deep CNN Architecture is shown in Fig. 11.

![Improved Deep CNN Architecture](image)

**Figure 11.** Improved Deep CNN Architecture

2.9.2. Result In this paper, the author was trying to analyze various feature extraction and classification techniques used for offline handwritten character recognition. Due to the high noise tolerance capability of ANN(Artificial Neural Network), it is used here to perform handwritten character recognition. They got excellent recognition results by using the above-proposed model.

3. Technical Comparison

Here we have given a technical comparison of different methodologies used for feature extraction and classification in TABLE I. The accuracy measures of classification of the above algorithms have been also given so that the researchers can select a particular algorithm with high performance. Among the different proposals that we have discussed in this paper, GBF-RLC with SQDF or MLP has given the highest performance.
Table 1. Technical Comparison

| Author             | Script     | Feature Extractor | Classifier | Accuracy |
|--------------------|------------|-------------------|------------|----------|
| Raju et al. [2014] | Malayalam  | GBF&RLC           | SQDF Or MLP| 99.66    |
| Shibaprasad et al. [2018] | Bangla   | CNN               | CNN        | 99.40    |
| Jino et al. [2017] | Malayalam  | Stacked LSTM      | LSTM       | 97.00    |
| Jino et al. [2019] | Malayalam  | 10 layer CNN      | SVM        | 96.90    |
| Chandrika et al. [2019] | Bangla   | 7 layer CNN       | CNN        | 96.80    |
| Durjoy et al. [2015] | Bangla   | 5 layer CNN       | SVM        | 95.60    |
| Binu et al. [2012] | Malayalam  | WE                | ELM        | 95.59    |
| Jomy et al. [2011] | Malayalam  | CCH               | NN         | 72.10    |

4. Conclusion

In this paper, handwritten recognition systems for characters and words from different languages are discussed. Various feature extraction and classification techniques are discussed here with deep learning platform. Big character set, similarity among characters and complex structures of compound characters are the major challenges that the researchers are facing in this area. We believe that our survey will be helpful for researchers in this field.

References

[1] Jino P J, Jomy John, Kannan Balakrishnan 2017 Offline Handwritten Malayalam character Recognition using stacked LSTM ICICICT-IEEE 17 1587
[2] Shibaprasad Sen 2018 Online Handwritten Bangla Character Recognition Using CNN: A Deep Learning Approach Springer Nature 695 413
[3] Binu P Chacko, VR Vimal Krishnan, G Raju, and P Babu Anto 2012 Handwritten character recognition using wavelet energy and extreme learning machine Springer 3(2) 149
[4] G Raju, Bindu S Moni, and Madhu S Nair 2014 A novel handwritten character recognition system using gradient based features and run length count Sadhana-Springer 39(6) 1333
[5] Jomy John, Pramod K. V, Kannan Balakrishnan 2011 Offline Handwritten Malayalam Character Recognition Based on Chain Code Histogram ICICT-IEEE 11 736
[6] Abhishek Hazra, Prakash Choudhary, Sanasam Inunganbi and Mainak Adhikari 2020 Bangla-Meitei Mayek scripts handwritten character recognition using Convolutional Neural Network Applied Intelligence- Springer Nature 20 1901
[7] PJ Jino, K Balakrishnan, U Bhattacharya 2019 Offline Handwritten Malayalam Word Recognition Using a Deep Architecture Soft Computing for Problem Solving-Springer Nature 816 913
[8] Durjoy Sen Maitra, Ujjwal Bhattacharya and Swapan K. Parui 2015 CNN Based Common Approach to Handwritten Character Recognition of Multiple Scripts ICDAR-IEEE 15 1021
[9] Chandrika Saha, Rahat Hossain Faisal and Md. Mostafijur Rahman 2019 Bangla Handwritten Digit Recognition Using an Improved Deep Convolutional Neural Network Architecture ECCE-IEEE 19
[10] Shreya Girdhar, Sachin Gupta, Bhaskar Kapoor 2018 Handwritten Character Recognition using Neural Networks: A Study of Various Feature based Classification Techniques IJSRD 6 2163
[11] Jomy John, K.V. Pramod, and B. Kannan 2011 Handwritten Character Recognition of South Indian Scripts: A Review National Conference on Indian Language Computing 1 11