Malayalam Handwritten Character Recognition Using Convolutional Neural Network

Pranav P Nair
Computer Science and Engineering
Govt. Engineering College
Thrissur, Kerala, India
pranav7nair@gmail.com

Ajay James
Computer Science and Engineering
Govt. Engineering College
Thrissur, Kerala, India
ajay@gectcr.ac.in

C Saravanan
Computer Centre
National Institute of Technology
Durgapur, West Bengal, India
cs@cc.nitdgp.ac.in

***Abstract*— Optical Character Recognition is the process of converting an input text image into a machine encoded format. Different methods are used in OCR for different languages. The main steps of optical character recognition are pre-processing, segmentation and recognition. Recognizing handwritten text is harder than recognizing printed text. Convolutional Neural Network has shown remarkable improvement in recognizing characters of other languages. But CNNs have not been implemented for Malayalam handwritten characters yet. The proposed system uses Convolutional neural network to extract features. This is method different from the conventional method that requires handcrafted features that needs to be used for finding features in the text. We have tested the network against a newly constructed dataset of six Malayalam characters. This is method different from the conventional method that requires handcrafted features that needs to be used for finding features in the text.**

***Keywords*— Feature Extraction, Classification, Machine recognition, Convolutional Neural Network, CNN, Malayalam.**

I. INTRODUCTION

Deep learning Techniques has achieved top class performance in pattern recognition tasks. These include image recognition [1, 2], human face recognition [3], human pose estimation [4] and character recognition [5, 6]. These deep learning techniques have proved to outperform traditional methods for pattern recognition. Deep learning enables automation of feature extraction task. Traditional methods involve feature engineering which is to be done manually. This task of crafting features is time consuming and not very efficient. The features ultimately determine the effectiveness of the system. Deep learning methods outshine traditional methods by automatic feature extraction.

Convolutional Neural Networks (CNN) is a popular deep learning method and is state of the art for image recognition. CNN has achieved a breakthrough in the IMGENET challenge 2011. The CNN used in the challenge was Alexnet and gave an error rate of 16% in comparison to 25% in 2010. From then on it was CNN all the way. CNN is very suitable to represent the image structure. The properties of CNN that makes this possible are the local connectivity strategy and the weight sharing strategy [7].

Handwritten character recognition is a difficult task as the characters usually has various appearances according to different writer, writing style and noise. Researchers have been trying to increase the accuracy rate by designing better features, using different classifiers and combination of different classifiers. These attempts however are limited when compared to CNN. CNNs can give better accuracy rates but it has some problems that needs to be addressed.

Malayalam is one among the twenty two scheduled languages in India and is the official language in the state of Kerala, where more than 95% people use Malayalam for communication [10]. Malayalam characters are complex due to their curved nature and there are characters which are formed by the combination of two characters. These along with the presence of 'chillu' make recognizing Malayalam characters a challenging task.

We attempt to use CNN to achieve better accuracy rate in Malayalam handwritten character recognition. The rest of the paper is organized as follows. Section 2 is literature survey; section 3 gives the methodology and section 4is the conclusion.

II. LITERATURE SURVEY

Shailesh Acharya et al [8] proposed a Deep Learning Based Large Scale Handwritten Devanagari Character Recognition that used convolutional neural network for classifying Devanagari handwritten characters. The employed data set increment and dropout layer in order to reduce over fitting. They tested two models of the network, model A consisted of three convolution layers and one fully connected layer and model B was a shallow network. The highest testing accuracy for Model A was 0.98471 and for model B was 0.982681.

Prashanth Vijayaraghavan et al. [9] proposed a handwritten character recognition system for Tamil characters using convolutional neural network. They augmented the

ConvNetJS library for learning features by using stochastic pooling, probabilistic weighted pooling, and local contrast normalization get an accuracy of 94.4% on the IWFHR-10 dataset. Anitha et al [10] proposed Multiple Classifier System for Offline Malayalam Character Recognition. The features used are the gradient and density based features. The best combination ensemble with an accuracy of 81.82% was reported by using the Product rule combination scheme.

G Raju et al. [11] 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 characters4. The authors have also applied gradient features for the recognition of Malayalam vowels in5. Arora et al. [12] proposed a multiple classifier system using chain code histogram and moment invariants for the recognition of Devanagari character recognition.

Rajashekararadhya S. V et al. [13] suggests an efficient approach for handwritten numeral recognition in Kannada and Tamil. Projection distance was used as feature and additionally zone based method was also used for accurate recognition of numeral. For training and testing, they used Nearest Neighbor classifier.

They divided the whole image into 25 equal parts and from each zone; pixel distance for grid column is computed from image centroid. If more than one pixel is found in a column of that grid, average pixel distance is computed and stored. Repeat this process for entire grid to get 250 features (10 features for a zone). An average of 95% accuracy was achieved.

Giridharan. R et al. [14] propose zoning method to retrieve information from temple Epigraphy. They decomposed the image in several ways. Decompositions to vertical zones, horizontal equal zones, right diagonals, left diagonals, octants, diagonal quadrants, quadrants etc were used as zoning. A total of 54 zones were obtained. In each zone, they calculated the density of black pixel. Perceptron was used for the recognition purpose. It takes multiple inputs and produces single output using a linear combination of input values. Another OCR was advanced by M Abdul Rahman and M S Rajasree [15] using wavelet transform feature extraction techniques and neural networks.

There is no OCR system that yields 100% accuracy. CNN was not yet implemented for handwritten character recognition in Malayalam. CNN has shown a great deal of improvement in the accuracy rates in other languages.

The paper is organized in the remaining portion as follows. In Section III discusses the proposed method. Section IV concludes the paper.

III. PROPOSED METHOD

There is no standard dataset for handwritten Malayalam characters. CNN requires a large set of training images. CNN achieves a high accuracy rate only if it is trained with a substantially large training set. This is one of the biggest challenges given the time period is short. However from literature survey, it is clear that there are techniques that can be applied to increase the number of dataset images. Overall architecture of the proposed system is shown in figure 1. The input is first scanned using a scanner or taken as a photograph using a smart phone. The kernel weights are initialized using Gaussian distribution.

The proposed method consists of the following stages:

A. Pre-processing

In the preprocessing stage, the character image is processed for removing all the undesirable entities from an image to make the process of recognizing easier. The input images are resized to a suitable format. It must not be too large or too small. If the image is too large, the amount of computation required will be high. If the image is too small, it will be difficult to fit it into a large network. Larger images are cropped and padding will be applied to smaller images to achieve a standard size. Padding is the process of adding white pixels to an image, which means that we are increasing the background of the image.

B. Dataset Creation

There is no open source dataset available for handwritten Malayalam characters. Hence it was necessary to build a new dataset from scratch. Creating a dataset is time consuming and requires a lot of effort. To start with, we decided to first build a dataset of the first six characters of Malayalam. Characters written by 112 different people were collected. A complete Malayalam dataset is being constructed. The complete dataset will have 3 times the variety of the current dataset.

C. Dataset Augmentation

A large dataset is required for training the CNN. In order to attain this, the images that are already obtained is modified and transformed to get a large number of variations. Affine transformation is a linear mapping method that preserves points, straight lines, and planes. Sets of parallel lines remain parallel after an affine transformation. Translation, Scaling, Sheering and Rotation are the four major affine transformation. Different translations are used to augment the dataset. Affine transformation is a linear mapping method that preserves points, straight lines, and planes. Sets of parallel lines remain parallel after an affine transformation. Gaussian smoothing is the result of blurring an image by a Gaussian function. The visual effect of this blurring technique is a smooth blur resembling that of viewing the image through a translucent screen. Salt-and-pepper noise is a form of noise sometimes seen on images. It presents itself as sparsely occurring white and black pixels. Contrast and brightness level of an image is changed. After data augmentation a dataset of nearly 2 lakh images will be obtained.

D. CNN Modelling

This is the most important step. CNN modeling means modeling the structure of CNN. The number of convolution layers, max pooling layers, ReLu layers and fully connected

layers needs to be chosen. It is not possible to determine the exact number of layers that will yield the best outcome. Hence it is vital to try several configurations of the network and choose which network best suits. Size of Feature map=m*n+1, n <= m, where m is the height and width of the image, n is the height and width of the convolution layer.

Max pooling has shown to be the most effective pooling strategy and hence will be used by our system. Back propagation using gradient decent will be used as the learning rule. The LeNet-5 was the most used for character recognition and hence it is the model to which we compare our new model. LeNet architecture is shown in figure 2 [18]. The number of neurons in each convolutional layer was adjusted to get the maximum accuracy rate. An increasing pattern will be followed the number of neurons for each layer. A dropout layer will be used to aid in the training process. The dropout layer decreases the complexity and training time of the network.

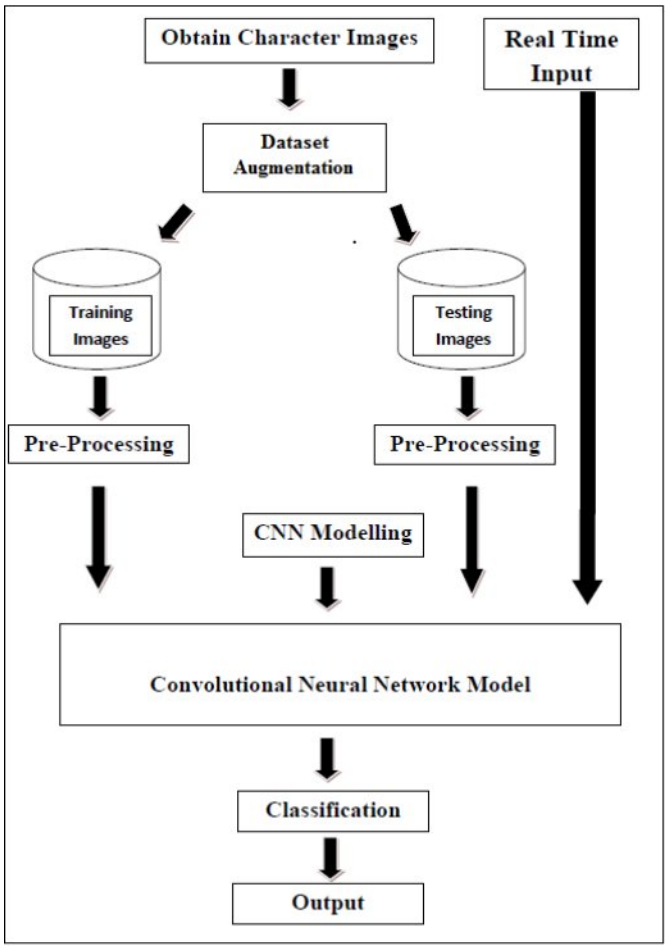

Figure 1: Overall system Architecture.

E. Classification

The final layer of the CNN is a Softmax layer and this softmax layer is used for classifying the given input image. This softmax layer is used to classify the character. The softmax function has a value between 0 and 1. The sum of output of all the classes also sums to 1. The class with the maximum value will be selected as the class for a particular input image.

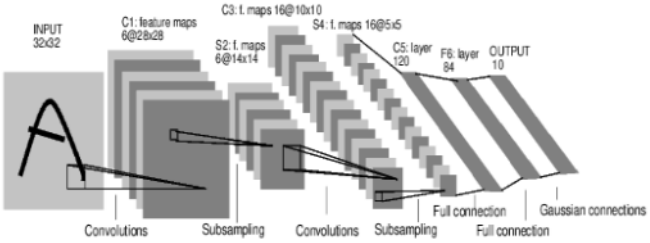

Figure 2: LeNet Architecture[18].

F. Testing

Testing module deals with the test images. Test images are obtained by splitting the augmented dataset randomly. It will first preprocess the input image and it will classify the unlabeled test data. Test data is not labelled in the sense it should be recognized by the machine. Labels are assigned to each of test images by the network and then the accuracy is measured. In the post processing stage of the proposed system, produced classifier output is mapped to the character Unicode. The output of the classifier will be some integer labels. This integer label should be converted into corresponding character Unicode. The Unicode is written in a text file.

IV. CONCLUSION

OCR has a wide variety of real time applications. It can be used for office automation. This work implements a handwritten Malayalam character recognition system. The proposed method uses CNN to extract and classify Malayalam characters. Both Sample generation and CNN modelling are time consuming tasks and the later also requires a CUDA enabled GPU for parallel processing.

Preprocessing helps to remove the undesired qualities of an image and hence can play an important role in increasing the role. So is the sample generation process that reduces overfitting. The drop out layer also reduces overfitting while also decreasing the overall training time. CNN has proved to be the state-of-the-art technique for other languages and hence provides the chance for giving higher accuracy rate for Malayalam characters too.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks", In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions". *arXiv preprint arX- iv:1409.4842*, 2014.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification", *IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1701–1708. IEEE, 2014.

[4] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler "Joint train- ing of a convolutional network and a graphical model for human pose estimation", In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[5] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis. In 2013 12th International Conference on Document Analysis and Recognition", volume 2, pages 958–958. IEEE Computer Society, 2003.

[6] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber. "Convolutional Neural Network Committees for Handwritten Character Classification", pages 1135–1139. IEEE, Sept. 2011.

[7] Li Chen, Song Wang, Wei Fan, Jun Sun, Satoshi Naoi, "Beyond Human Recognition: A CNN-Based Framework for Handwritten Character Recognition",3rd IAPR Asian Conference on Pattern Recognition, 2015.

[8] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks". *arXiv preprint arX- iv:1312.6082*, 2013.

[9] Shailesh Acharya, Ashok Kumar Pant, Prashnna Kumar Gyawali "Deep Learning Based Large Scale Handwritten Devanagari Character Recognition", 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2015.

[10] Prashanth Vijayaraghavan and Misha SraHandwritten "Tamil Recognition using a Convolutional Neural Network", MIT Media Lab.

[11] Anitha Mary M.O. Chackoa, Dhanya P.M, "Multiple Classifier System for Offline Malayalam Character Recognition", *International Conference on Information and Communication Technologies (ICICT)*, 2014.

[12] 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.

[13] 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.

[14] Abdul Rahiman M and Rajasree M S, "An Efficient Character Recognition System for Handwritten Malayalam Characters Based on Intensity Variations", *International Journal of Computer Theory and Engineering, Vol. 3, No. 3*, June 2011.

[15] Panyam Narahari Sastry, T.R. Vijaya Lakshmi, N.V. Koteswara Rao,T.V. Rajinikanth, Abdul Wahab, "Telugu handwritten character recognition using zoning features",IEEE, 2014.

[16] Manoj Kumar Mahto, Karamjit Bhatia R. K. Sharma, "Combined horizontal and vertical projection feature extraction technique for Gurmukhi handwritten character recognition", *International Conference on Advances in Computer Engineering and Applications (ICACEA)*, 2015.

[17] Manju Manuel and Saidas S. R, "Handwritten Malayalam Character Recog nition using Curvelet Transform and ANN", *International Journal of Computer Applications (0975 8887) Volume 121 No.6,* July 2015.

[18] Yann LeCun, Leon Bottou, Yoshua Bengio and Patric Haffner, "Gradient-Based Learning Applied to Document Recognition", PROC. OF THE IEEE, NOVEMBER 1998.

.