International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 3, pp: 122-125 www.ijaem.net ISSN: 2395-5252

Malayalam Handwritten Character Recognition

Athullya Bhaskar K R, Rameez Mohammed A
Computer Science and Engineering GEC Palakkad Palakkad, India
Computer Science and Engineering GEC Palakkad Palakkad, India
Corresponding author:Athullya Bhaskar K R

Date of Submission: 15-07-2020 Date of Acceptance: 31-07-2020

ABSTRACT—Recognizing handwritten text is harder than recog- nizing printed text. Convolutional Neural Network has shown remarkable improvement in recognizing characters of other languages. 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. This paper proposes a CNN architecture for classification of handwritten characters in Malayalam language. 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.

Index Terms—Classification, Convolutional Neural Network, Dataset Augmentation, Segmentation

I. EXISTING SYSTEM

Malayalam Script digitalization is a common requirement in areas of literature publications, for film screenwriters, news reporters, writers etc. It has a number of other uses also. In the existing system the handwritten documents has to be explicitly converted into digital form by reading it with human eye and type it manually to a computer. This is a time consuming and a tedious task which require lot of manual effort to complete the intended exercise. With the support of a smart character recognition mechanism, almost all hindrances can be eradicated.

II. INTRODUCTION

Optical character recognition is to reduce tedious manual work of converting images containing characters to texts for recent decades. Optical character recognition has been success- fully implemented in many areas which greatly reduce manual work of encoding physical document to machine encoded format. Different methods are used in OCR for different languages such as Bayesian theory, Hidden Markov Model, Template Matching and Neural Networks. The recognition task of handwritten characters is rather complex and is a great challenge to researchers as the solution should be able to cope with the challenges in identifying the characters from a variety of writing styles, slants of personal interest. Especially in a language like Malayalam which is having a complex structure and very identical nature of character set. This project aims to develop a system to recognize handwritten Malayalam characters using Convolutional Neural Network (CNN) which is a very popular deep learning technique. Malayalam characters are having a curved nature and there are many glyphs which have very similar looks and some characters are the combination of other characters. All these difficulties makes Malayalam characters are hard to detect as it requires deep machine learning models to classify every characters of it. It needs better deep learning models to correctly classify the characters in the language. An interface should be developed which enables the user to input an image of the handwritten Malayalam script and do proper preprocessing segmentation and give the Unicode Malayalam characters as output.

III. METHODOLOGY

A. Deep Learning

Deep Learning is a subset of machine learning concerned with algorithms inspired by the structure and function of the brain called Artificial neural networks (ANN). In deep learning we don't need to do handcrafted feature extraction like we do in classifiers like SVM and Random Forest. The ANN extracts features by its own. For character classification, this project uses Convolutional neural network (CNN) which is a very popular neural network for finding patterns in data.

B. Convolutional Neural Networks

Convolution neural network algorithm is a special kind of feed forward multilayer perceptron which is basically an Artificial Neural Network (ANN). It has proven its design for identification of patterns from two dimensional data. It has been

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 3, pp: 122-125 www.ijaem.net ISSN: 2395-5252

applying in face recognition, image classification, natural language processing etc. The main difference between a CNN and an ANN is that CNN uses parameter sharing which makes the computation a lot more easier. A typical CNN Always has one input layer, convolution layers, pooling layers, ReLU(Rectified Linear Unit) layers, fully connected layers and one output layer. In addition to this, the number of layer in a CNN is subjected to change according to the classification requirements. Convolutional netwoks use different multilayer perceptrons designed to require minimal preprocessing. Con- volution operation is applied to the input of convolutional layer, then passing the result to the next layer. Convolutional networks include local or global pooling layers. It combines the outputs at one layer into a single neuron in the next layer. This is done for down sampling a feature map obtained as the output of a convolutional layer. The final fully connected

layers connects every neuron in the previous layer to the next layer. So the final downsampled featuremap is stretched into a single vector and is given as input to the fully connected layer. The output of the final fully connected layer contains the required number of classes.

IV. PROPOSED SYSTEM

The product can be effectively used in situations where we need to digitalize a Malayalam handwritten script , by using an intelligent system to recognize the handwritten characters and make a digital equivalent of the scripts we can greatly reduce the manual effort of character recognition and data entry for digitalizing a handwritten document .Malayalam Script digitalization are a common requirement in areas of literature publications, for film screenwriters , news reporters etc. With a support of a smart character segmentation mechanism, the system can be used to meet the requirements of all these areas. The overall system architecture is shown in Fig.1. The system mainly consists of four phases. They are:

1) Dataset creation and augmentation
2) Creating a CNN model
3) Training the CNN model
4) Deploy the model

A. Dataset Creation and Augmentation

Creating a dataset is time consuming and requires a lot of effort. For this project the dataset was collected from a private organization and modified. The dataset consisted images of 48 Malayalam characters. A very 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. Fig.2 shows the overall flow of this process. Affine transformation is a geometric transformation that preserves lines and parallelism. It is a linear mapping method that preserves points, straight lines, and planes. After an affine transformation, the sets of parallel lines remain parallel. Different translations are used to augment the dataset . The result of blurring an image by a Gaussian function is Gaussian smoothing . A form of noise sometimes seen on images are called Salt-and-pepper noise. It presents itself as scarse occurring white and black pixels. Contrast and brightness level of an image is changed.

B. Creating a CNN model

Convolutional neural network layer types mainly include three types, namely Convolutional layer, pooling layer and fully connected layer apart from the input and output layer. There are many standard CNN architectures available which are proven their classification capabilities in many tasks ,eg. LeNet , Alexnet, VGGNet, Inception etc. all these architecture differ due to the difference in number of hyper parameters of the network such as number of convolution layers, pooling layers ,fully connected layers ,the number of filters used in each layers, dropout rate at each layer, L2 or L1 regularization parameters, activation function type (ReLU, Sigmoid, Tanh etc.). The input size of images dataset is fixed to 86x86, so

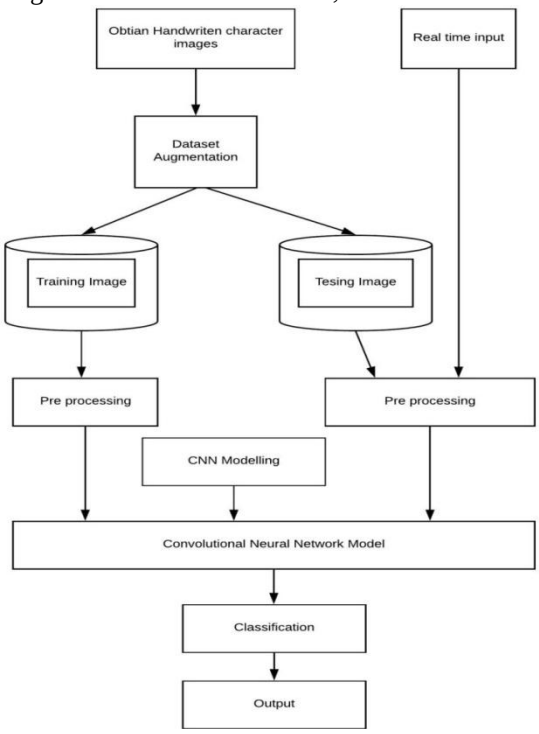

Fig. 1. System Architecture

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 3, pp: 122-125 www.ijaem.net ISSN: 2395-5252

the initially the network is designed to occupy tensors of shape [86,86].This model uses 3x3 filters for convolution,2x2 filters for max-pooling and ReLU(Rectified Linear Unit) as activation function in Non linearity layer. Batch normalization is applied to get better results, Adam is used as optimizer for gradient descend algorithm to perform. During training a single image of a Malayalam character with one-hot encoded label passes through all the layers and according to the applied gradient descent strategy here, the weights are updated. The output layer contains 48 classes each represents a character in dataset.

C. Training the CNN model

The model has to be trained with the dataset created. For this we need to prepare the dataset to input and labels format so that the model understands it. The dataset is processed and a list is created which consists of numpy array of images along

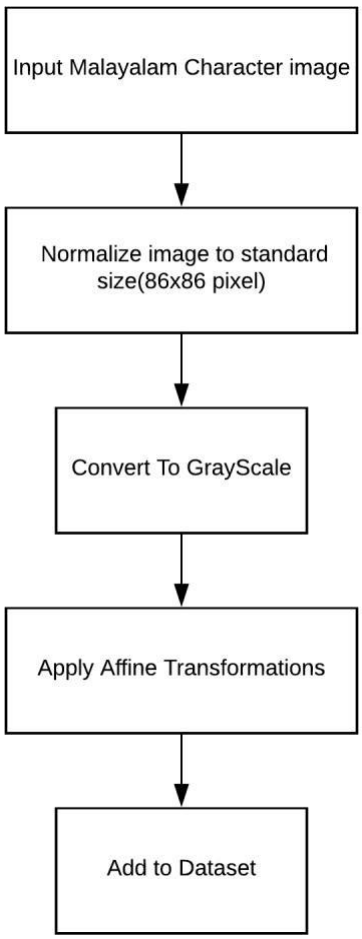

Fig. 2. Dataset Creation and Augmentation

with the one hot encoded label. The preprocessed dataset is divided into training set and testing set . There were above 70000 images for training and above 20000 images for testing. Validation set used here to test during the training time itself so that we can see whether the model overfitts or not during each epoch. To reduce the complexities of memory inefficiencies while training the entire dataset is divided into batches and each batch is given to the network for training.

D. Deploy the model

After successful training, the model has to be deployed so that the user can input an image of handwritten Malayalam script and the system predicts the Unicode Malayalam characters as output using saved CNN Model. First the image of Malayalam handwritten character has to be read. And then applying proper segmentation mechanisms, the words are separated and then characters are separated. Each character that is segmented are given to model and prediction is made. For each predicted class the character is mapped to corresponding Malayalam Unicode character. The Segmentation algorithm is as follows:

1) Read image
2) Convert image into grayscale
3) Apply morphologyEx operation to get the words as a single contour
4) Store the contour bounding boxes as a list
5) For each bounding box , apply morphologyEx operation with structuring element of small kernel size
6) Find contours and draw bounding boxes
7) Sort the bounding boxes in the increasing order of x axis
8) Crop each bounding box coordinate from input image and store it in a list

V. RESULTS AND DISCUSSIONS

This project uses Convolutional neural network to classify Malayalam handwritten characters. For that a CNN model is created and the hyper parameters are tuned to get the increased accuracy. The model gave a training accuracy of 97.02% with loss 0.3718, validation accuracy of 96.18% with loss 0.3242 and a testing Accuracy of 96.35%.

VI. CONCLUSION

Handwritten character recognition is a difficult task as the characters usually has various appearances according to different writer, writing style and noise. Malayalam characters are complex due to their curved nature. Here this project built a system that recognizes Malayalam handwritten

characters using Convolutional neural network approach. CNNs can give better accuracy rates. This project will greatly help to reduce the digitalization work of any Malayalam handwritten document. Both Sample generation and CNN modeling 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. So this 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.

VII. FUTURE ENHANCEMENTS

This model can classify 48 Malayalam characters .There are more symbols in Malayalam language .It has to be improved to classify all the characters and symbols that is present in Malayalam language. Malayalam language has a peculiarity that unlike English language a character may have different meaning with respect to its position. In current system , if the input image containing words that are connected the current system may not give intended results. Some more researches have to be done on this. The state based model such as RNN (Recurrent neural networks) can be incorporated with this model so that it can improve the current model.

ACKNOWLEDGMENT

First and foremost I wish to express my whole-hearted indebtedness to God Almighty for his gracious constant care and blessings showered over me for the successful completion of the project. Moreover, I express my deep gratitude to family, friends and teachers for their whole hearted support.

REFERENCES

[1]. N. P. Kishna Thulasi. 2017.”Intelligent tool for Malayalam cursive handwritten character recognition using artificial neural network and Hidden Markov Model”, In Proceeding International Conference on Inventive Computing and Informatics (ICICI), IEEE.

[2]. P. Nair Pranav , Ajay James, and C. Saravanan. 2017. ”Malayalam handwritten character recognition using convolutional neural network”, In Proceeding International conference on Inventive Communication and Computational Technologies (ICICCT), IEEE.

[3]. Vijayaraghavan , Prashanth , and Misha Sra. 2014. ”Handwritten tamil recognition using a convolutional neural network”, In Proceeding Inter- national Conference on Information, Communication, Engineering and Technology (ICICET).

[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 39.6 (2014): 1333-1355.

[5]. Rahiman Abdul, and M. S. Rajasree.2011. ”An efficient character recognition system for handwritten Malayalam characters based on intensity variations”, In Proceeding International Journal of Computer Theory and Engineering 3.3 (2011): 369.