DIGITALIZATION OF TAMIL HANDWRITTEN CHARACTERS RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS(CNN)

Ram Kumar S[1], Sivamurugan A[2], Sai Vignesh M[3], Shanmugam K[4]

[1] *Student, Computer Science and Engineering, SRM Valliammai Engineering College, Tamilnadu, India*
[2] *Student, Computer Science and Engineering, SRM Valliammai Engineering College, Tamilnadu, India*
[3] *Student, Computer Science and Engineering, SRM Valliammai Engineering College, Tamilnadu, India*
[4]*Assistant Professor, Computer Science and Engineering, SRM Valliammai Engineering College, Tamilnadu, India*

ABSTRACT

Now-a-days digitalization becomes an important one for documents preservation. Some Tamil Handwritten characters need preservation, like land documents etc. So we try to overcome the difficulty of paper preservation by digitalizing it. The aim of this Project is to require Handwritten set of Tamil Characters as input within the format of image to process the character, train the Convolution Neural Network algorithm to acknowledge the pattern and convert the recognized characters to a Printed document. Convolutional Neural Network then attempts to work out if the computer file matches a pattern that the Neural Network has memorized. Optical Character Recognition deals with a crucial concern issue of handwritten character classification. To beat the difficulty of knowledge recognition among similarities, Convolutional Neural Network will provide more accuracy of character recognition. Convolutional Neural Networks (CNN) are playing an important role nowadays in every aspect of computer vision applications. The art of CNN is used in recognizing Tamil handwriten characters in offline mode. CNNs differ from traditional approach of Tamil Handwritten Character Recognition (THCR) in extracting the features by the methods of preprocessing ,normalization, feature extraction and classification .we've developed a CNN model from scratch by training the model with the Tamil characters in offline mode and We have achieved good accuracy results on obtained datasets. This work is for digitalizing offline THCR using deep learning technique.

Keyword: *Preprocessing,Normalization, Feature extraction,Convolutional Neural Network(CNN),Digitalization.*

.

1. INTRODUCTION

Tamil is one of the standard Indian languages which is predominantly utilized in Southern India, Srilanka and Malaysia. The speciality of Tamil language is every sound pronounced features a syllable in Tamil. The economy of characters to represent a word is minimal in Tamil language. The tiniest unit of Tamil script is syllable. These syllabic units of Tamil script has 12 vowels, 18 consonants and a special character Ayudha Ezhuthu(ஃ). There are 247 characters, among that 206 compound characters are formed by vowels and consonants. There are about 5 borrowed consonants from Sanskrit, when these sanskrit consonants combined with tamil vowels would yield another 60 compound characters so on make a filled with 307 characters. The entire Tamil listing could even be represented by combinations of 156 distinct characters.For an example, கௌ (pronounced as kow) which is represented by combining ’,’க’ and ’ள’. Technology has always played a very important role in enhancing various arenas. Integration of many technologies has always proved to be boon in every aspect of studies. The scope of Tamil language research studies enhances with the mixture of technology. Being one in every of the oldest classical language Tamil deserves to be simplified and reach to everyone who desire to travel looking out it and use it. Tamil language has the pride status of 15th most language within the globe. Technology integration in Tamil language facilitate researchers and users to explore the language further.

1.1 LITERATURE SURVEY

In the existing model, a typical approach for THCR using traditional machine learning techniques would follow preprocessing, feature extraction, classification and then predicting the new characters. Shanthi and Duraiswamy [2] proposed a model , whose features extracted are pixel densities and SVM classifier is used for classification process of 106 classes. They achieved an accuracy of 82.04% for 34 characters. Jose and Wahi [3] used wavelet transform method for feature extraction and for classification used a Backpropagation neural network with which have achieved 89% recognition accuracy in it. Sureshkumar and Ravichandran [4] have extracted features from each character glyphs with various attributes and classified using Support Vector Machines (SVM), Self Organizing Maps (SOM), Fuzzy network, RCS algorithm and Radial basis function. Bhattacharya et al[5] have proposed a two stage recognition method in which an unsupervised clustering is used in first stage for grouping the character classes and a supervised classifier is used in second stage for recognizing the characters, thereby an accuracy of 89.66% was achieved.

2. PROPOSED SYSTEM

In the proposed system The 216 consonant vowels called as composite characters(modified) are recognized along with 12 vowels and 18 consonants.Proposed technique – Convolutional Neural Network which provides promising results with greater accuracy and less time consumption. Technique Definition- Convolutional Neural networks are particularly useful for solving problems that cannot be expressed as a series of steps, such as recognizing patterns, classifying them into groups, series prediction and data mining. Convolutional neural network then attempts to determine if the input data matches a pattern that the neural network has memorized and also it gives in a digital text format.

The dataset of 70 tamil characters are choosen from hp-laps-tamil-datset and for every character there are minimum 200-250 samples are taken and it is trained by convolutional neural networks and its accuracy of training and testing are obtained .This model provides better access of handwritten datasets than previous model of component labeling and OCR techniques.Here the dataset of emnist is taken it has digits and English handwriiten characters so we tried a new approach to predict both numbers, English and Tamil characters. Our major part is on newly added hp-labs tamil dataset is normalized to emnist dataset. Emnist dataset is a model of online dataset which a specified class labels with it ,so we added 70 more tamil characters with it.

3. ARCHITECTURAL DESIGN

The character is recognized by series of process steps as it undergoes pre-processing ,feature extraction, normalization and classification. This project majorly concentrates on classification part as Convolutional Neural Networks is used for classification method. CNN are the widely used deep learning models in handling image related tasks like image recognition, image classification, image captioning etc. These networks are generally a combination of convolution layers, pooling layers and fully connected layers.

3.1 PREPROCESSING OF IMAGE

Image preprocessing steps RGB image converted into gray scale image using NTSC gray scale conversion which is usually accustomed convert RGB to gray scale conversion. To remove ,the noise we have use median filter.The salt and pepper noise which introduce during the image acquisition is deduct using this filter.Apply 3 x 3 square median filtering technique to remove noise.Normalize the acquire image by converting it to 28 × 28 pixels.This size gives enough information of the image when the processing time is low.The image is converted from Gray scale to binary image that is an image with pixels 0‟s (white) and 1‟s (black).This conversion can take place because it conveys proper information of Word Structure.Converting into binary image,removing unnecessary pixel values.

3.2 FEATURE EXTRACTION

Feature extraction involves reducing the amount of resources required to explain an outsized set of knowledge . When performing analysis of complex data one among the main problems stems from the amount of variables involved. Analysis with an outsized number of variables generally requires an outsized amount of memory and computation power, also it's going to cause a classification algorithm to overfit to training samples and generalize

poorly to new samples. Feature extraction method is of constructing combinations of the variables to urge around these problems while still describing the info with sufficient accuracy.

3.3 NORMALIZATION

Normalizing the images by changing the pixel intensities to find the plotting area. Thus, the plotting of points is noted and its intensities are obtained by this normalizing method. In machine learning , normalizing is by providing a set of functions to normalize the image into 28X28 pixels.

3.4 CLASSIFICATION

CNNs are the widely used deep learning models in handling image related tasks like image recognition, image classification, image captioning etc. These networks are generally a combination of convolution layers, pooling layers and fully connected layers . These three blocks are used to construct a CNN model by varying the number of blocks, adding or deleting a block. Various architectures have been developed since 2012 Imagenet competition which had reduced the misclassification rate from 15.6% to 3.7% .

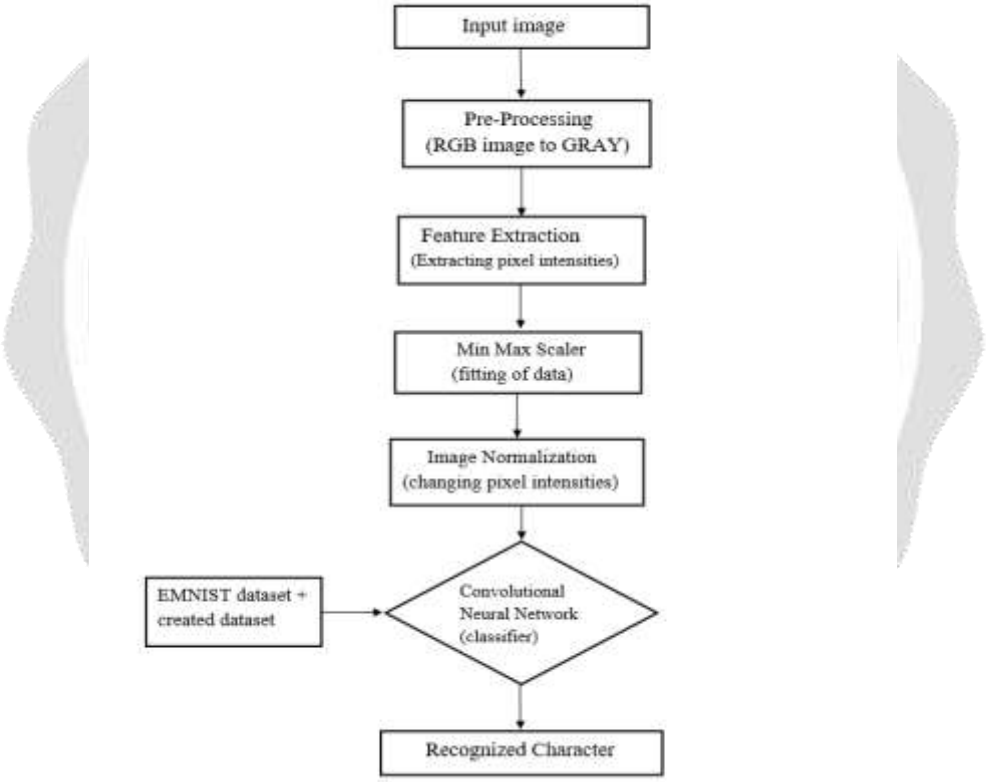

Fig -1: Process Flow Diagram

4. BACKGROUND WORK(CNN ALGORITHM)

CNNs are the widely used deep learning models in handling image related tasks like image recognition, image classification, image captioning etc. These networks are generally a combination of convolution layers, pooling layers and fully connected layers .Convolutional Neural Network is best for image classification. As the input of Handwritten characters are given as image inputs then it undergoes pre-process and normalization techniques ,finally it move on to classification part . Convolutional Neural Networks trains the images and stored as a model. This model checks the given test images to recognize the character.

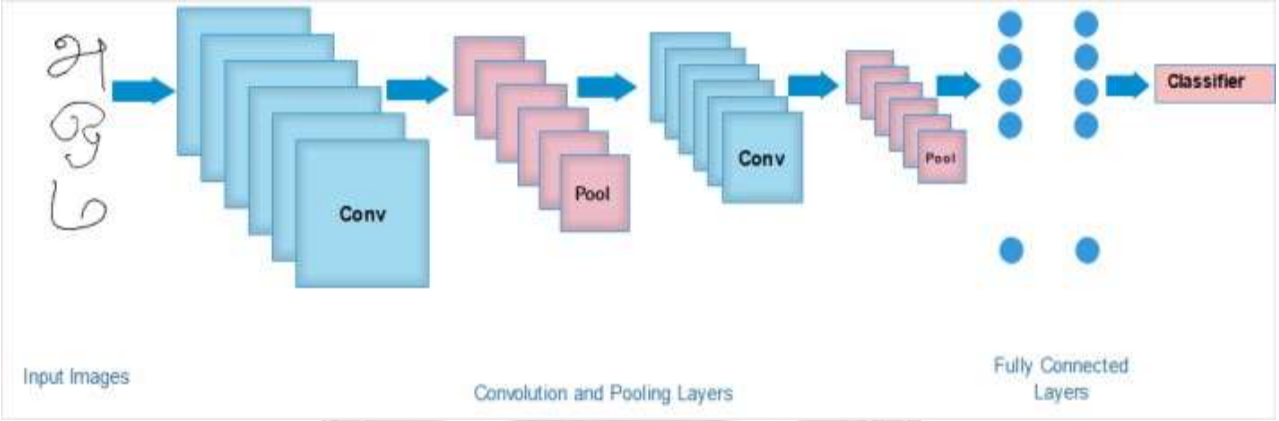

Fig -2: CNN Architecture diagram

4.1 CONVOLUTIONAL LAYER:

Convolution layer differ with a neural network in a very way that, not every pixel (a neuron) is connected to the subsequent layer with a weight and bias, but the complete image is split as small regions (say a nxn matrix) and weights and bias are applied over it. These weights and bias are remarked as filters or kernels which when convoluted with every small region within the input image would yield feature maps. These filters are the easy 'features' that's searched within the input image within the convolution layer. The amount of parameters required for this convolution operation would be minimal because the same filter is traversed over the complete image for one feature. The amount of filters, size of the local region, stride, and padding are the hyper parameters of convolution layer. supported the dimensions and genre of the input image, these hyper parameters can be tuned to attain better results.

4.2 POOLING LAYER

In order to cut back the spatial dimension of the image similarly because the number of parameters, thereby to cut back the computation, this pooling layer is employed. This layer performs a set function over the input, hence no parameters are introduced. differing types of pooling are available like average pooling, stochastic pooling, max pooling. Max pooling is that the most ordinarily used pooling algorithm, within which an nxn window is slid across and down the input with a stride values and for every position the most value within the nxn region is taken, thereby reducing the dimensions of the input. This layer provides translational invariance such even with a small variation within the position would still be able to recognize the image. But the situation information is lost because the size is reduced.

4.3 FULLY CONNECTED LAYER

In this layer the flattened output of the last pooling layer is fed as input to a totally connected layer. This layer behaves sort of a traditional neural network layer where every neuron of the previous layer is connected to this layer. Hence, the quantity of parameters during this layer are higher compared to the convolution layer. This fully connected layer is connected to an output layer which is mostly a classifier.

4.4 ACTIVATION FUNCTION

Different activation functions have been used across various architectures of convolution neural networks.Nonlinear activation functions such as ReLU, LReLU, PReLU, and Swish have proven better results when compared to the classic sigmoid or tangent functions. These nonlinear functions have helped in speeding up the training. In this work we have tried different activation functions and found ReLU to be more effective than others.

5. EXPERIMETAL RESULTS

The input image is given and its shown in fig-3 ,here the given image is shown as its real size of x and y axis are mentioned on it.

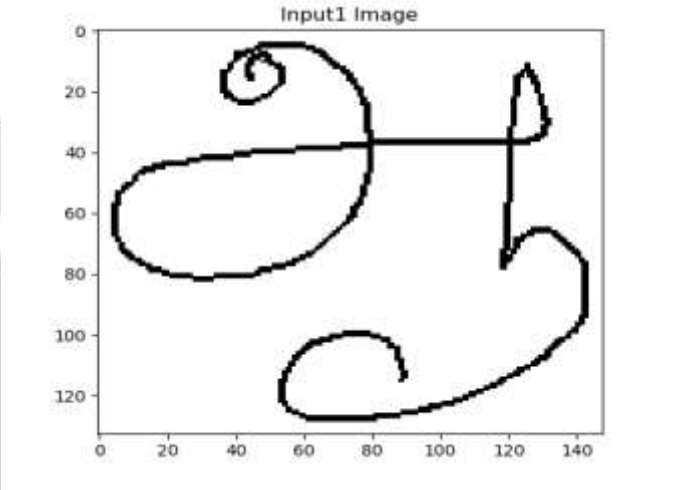

Fig -3: Input Image

The given input image is rescaled into 28X28 size of pixels and then its greyscaled from RGB colors and the image is shown in Fig-4 .

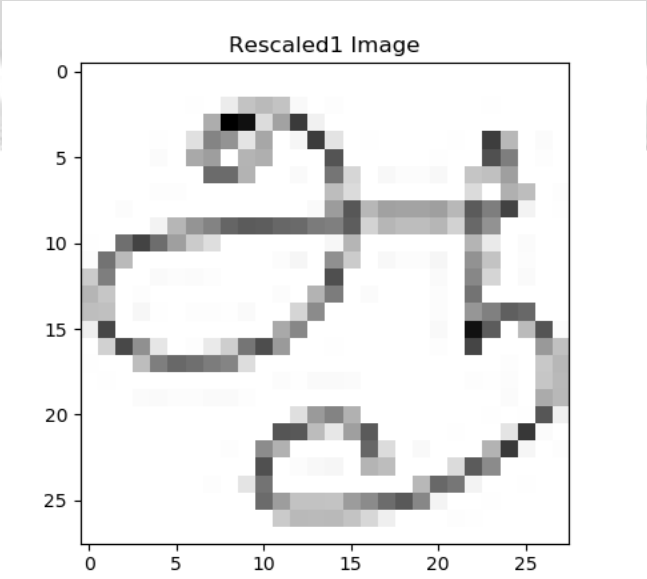

Fig -4: Rescaled Image

After the image is rescaled ,the normalization of image process is took place and its shown in Fig-5. It then,plot the pixels which has high threshold and it stored to check with the CNN trained model.

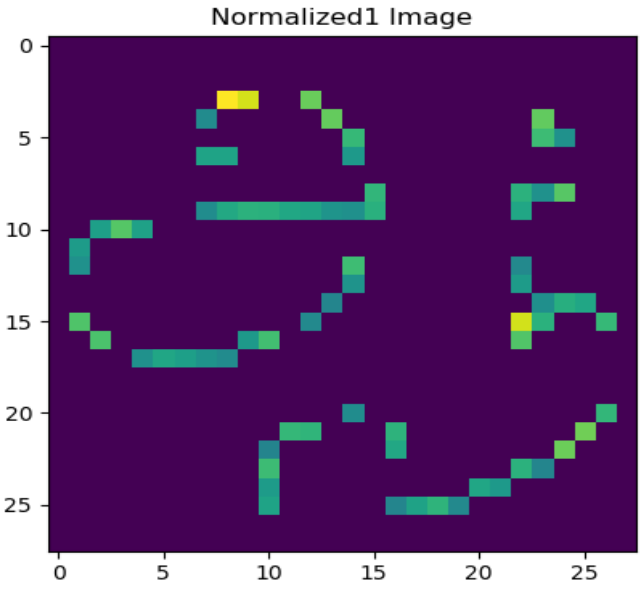

Fig -5: Normalized Image

6. CONCLUSION

In this project we recognized handwritten tamil characters ,by giving input image and then it is converted to a digital text format. This digital text can be stored in a document file ,and also a training and test accuracy of 93% is obtained by this CNN model which is better than Clustering and groupwise classification, support vector machine,Component labeling method and ANN.

7. REFERENCES

[1]. X. Zhang, F. Yin, Y. Zhang, C. Liu and Y. Bengio, "Drawing and Recognizing Chinese Characters with Recurrent Neural Network," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 849-862, 1 April 2018.

[2]. Shanthi, N., Duraiswamy, K., 2010. A novel SVM-based handwritten Tamil character recognition system. Pattern Anal. Appl. 13 (2), 173–180.

[3]. Jose, T.M., Wahi, A., 2013. Recognition of Tamil Handwritten Characters using Daubechies Wavelet Transforms and Feed-Forward Backpropagation Network. Int. J. Comput. Appl. 64 (8).

[4]. Sureshkumar, C., Ravichandran, T., 2010. Handwritten Tamil character recognition and conversion using neural network. Int. J. Comput. Sci. Eng. 2 (7), 2261–2267.

[5]. Bhattacharya, U., Ghosh, S.K., Parui, S., 2007. A two stage recognition scheme for handwritten Tamil characters. In: 2007. ICDAR 2007. Ninth International Conference on Document Analysis and Recognition. IEEE, pp. 511–515.

[6]. M. A. Pragathi, K. Priyadarshini, S. Saveetha, A. S. Banu and K. O. Mohammed Aarif, "Handwritten Tamil Character Recognition UsingDeep Learning," *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)* Vellore,India,2019,pp.1-5.doi:10.1109/ViTECoN.2019.8899614.

[7]. R. Kavitha and C. Srimathi, Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.004

[8]. Daniel Keysers, Thomas Deselaers, Henry A. Rowley, Li-Lun Wang, and Victor Carbune "MULTI-LANGUAGE ONLINE HANDWRITING RECOGNITION" IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, June 2017.

[9]. Muhammad Naeem Ayyaz, Imran Javed, Waqar Mahmood, "Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction", 2012, Pakistan journal of Engineering and Application Science,Vol. 10, pp. 57- 67.

[10]. Pal U., Sharma N., Wakabayashi T. and Kimura F., "Off-Line Handwritten Character Recognition of Tamil Script", In Proc. 9th International Conference on Document Analysis & Recognition, 2007, pp. 496-500.