Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 6, June 2020, pg.114 – 119

TAMIL-BRAHMI SCRIPT CHARACTER RECOGNITION SYSTEM USING DEEP LEARNING TECHNIQUE

Subadivya S
Dept. of Computer Science and Engg.
Rajalakshmi Engineering College, Chennai
subadivya.s.2016.cse@rajalakshmi.edu.in

Vigneswari J
Dept. of Computer Science and Engg.
Rajalakshmi Engineering College, Chennai
vigneswari.j.2016.cse@rajalakshmi.edu.in

Yaminie M
Dept. of Computer Science and Engg.
Rajalakshmi Engineering College, Chennai
yaminie.m.2016.cse@rajalakshmi.edu.in

Diviya M
Assistant Professor
Dept. of Computer Science and Engg.
Rajalakshmi Engineering College, Chennai
diviya.m@rajalakshmi.edu.in

Abstract – Character recognition is the process of detecting and processing a character or words and storing it in textual format. This project aims to build a character recognition system to digitize the ancient Tamil script and additionally translate the obtained script to modern Tamil. Although technologies around image processing are almost at perfection for languages like English and other modern languages, recognition of ancient scripts is still elusive. Conventional methods used by the Archaeological department convert the documents manually, which is possible only with the help of an expert. The proposed system will cross out such a need and will enable the common man to decipher the ancient scripts from inscriptions and manuscripts and also digitize, and also enable to process the data using Natural Language Processing. The system incorporates Convoluted Neural Networks to extract the features from each character to detect and translate the character to modern Tamil with 94.6 % accuracy.

Keywords — Character recognition, Natural Language Processing, Tamil-Brahmi, Deep Learning, CNN.

I. INTRODUCTION

The Brahmi script has numerous variants among which Tamil Brahmi was used to write the ancient Tamil inscriptions. The existence of Tamil-Brahmi has been dated between 300 BC and 100 AD. The Thirukkural, one of the greatest works on ethics and morality was written in Tamil Brahmi. The Thirukkural has been honored and praised by great intellects around the world. Our ancient scripts are the windows to our cultural heritage, civilizations and the origin of our society. The methods of interpreting these scripts involves an expert archaeologist who translates it into handwritten notes.

II. PROBLEM STATEMENT

The ancient scripts in inscriptions and manuscripts hold the key to Indian culture, heritage, ancient, scientific, mathematical advancements, like the Vedas and the shastras.

Decipherment of ancient Tamil inscriptions still involves conventional techniques and manual interpretation, which is time consuming, prone to human error and tedious.

The existing handwriting recognition systems face problems with the varying strokes, style and size of characters and tries to improve the accuracy of the systems.

III. PROPOSED SYSTEM

Our Tamil Brahmi Character recognizer aims to:

- To build an efficient system to recognize Tamil-Brahmi Script.
- To create a rich dataset in Tamil Brahmi Characters to train the model.
- Obtain an accurate interpretation of Tamil Brahmi characters to Modern Tamil.

IV. METHODOLOGY

In this system, the Convolutional Neural Network is used among other image processing techniques as it obtains high accuracy in similar use cases. The model is implemented using Keras library with TensorFlow in the backend. The data is created manually because of the difficulty in finding the dataset. Training images undergo a series of data preprocessing steps such as cropping, normalization and dimensionality reduction and then it is served to the network. The network is to be trained until an accuracy close to 100% is reached. Flask framework is used in order to make an application with this pretrained model which serves the end user with the modern Tamil character output when a Tamil-Brahmi letter is uploaded to the system.

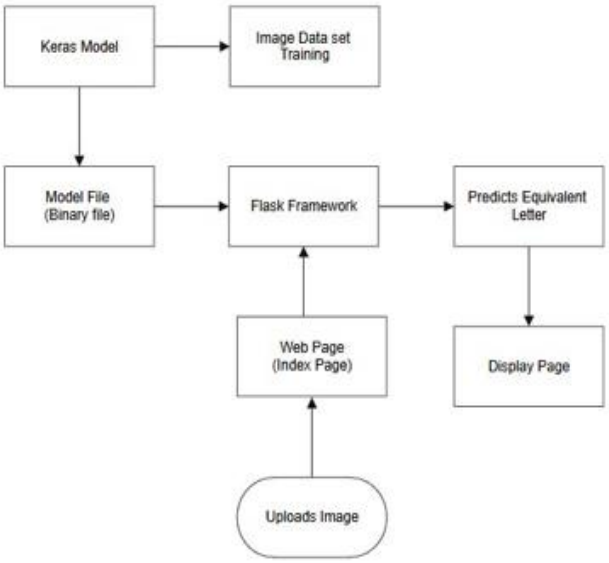

Fig.1 The above figure is the architecture diagram of the proposed model.

V. DATA PREPROCESSING

The dataset for this system is in the form of images, hence it is necessary to preprocess the image before using it in the network. As the dataset is manually created there could be complexity while processing it, so as to reduce the complexity it is necessary for the images to undergo certain data preprocessing steps in order to obtain utmost accuracy. In this data, only characters from early Tamil Brahmi characters are taken for processing. Thus, the use of pre-processing techniques may enhance a document image preparing it for the next stage in the character recognition system. Methods of data preprocessing include - A) Data Collection B) Cropping D) Normalizing Image Inputs E) Dimensionality Reduction F) Data Augmentation

A. DATA COLLECTION

As the dataset for this system is very old and ancient it is nearly impossible to collect thus a rich dataset for the characters in Tamil-Brahmi, which is manually created and fed to the training network. The created dataset is divided into two categories - train and test. The images in the training set is used to train the neural networks, whereas the images in the testing set is used to test the model for accuracy.

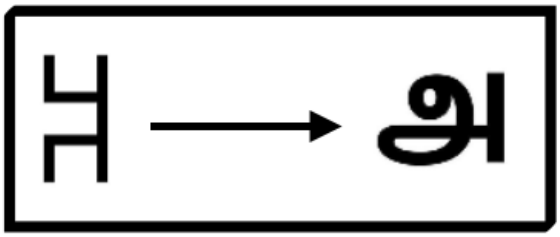

Fig.2 The above figure shows the translation of the letter 'aa' in Tamil Brahmi and modern Tamil.

B. CROPPING

The training images set is input in a specific size for the network to learn. Every input image is cropped to a certain height*width size. This helps the network to train the model more efficiently.

C. NORMALISING IMAGE INPUTS

The image set undergoes the normalization process as normalization transforms an n-dimensional grayscale image with intensity values in the range into a new image. To improve learning and recognition speed, for the high volume of data, down-sampling is employed by merging surrounding pixels into blocks. Finally, a 2D image is serialized into an array of blocks to conduct learning and recognition.

D. DIMENSIONALITY REDUCTION

Dimensionality Reduction is performed on the sample training data in order to remove features which are redundant and irrelevant and to prevent loss of information. The high dimensionality image is thus transformed into low dimensionality image with minimal loss of information.

E. DATA AUGMENTATION

Due to the lack of Tamil-Brahmi image samples, data augmentation methods such as shearing, rotation is performed over the available training set images to generate more training samples. This enhances the model by improving efficiency and achieving higher accuracy.

VI. CONVOLUTION LAYER AND FEATURE EXTRACTION

The model takes a Tamil-Brahmi input image, processes it and classifies it under its respective modern Tamil character. The model is trained using a set of images. Each image will pass through a series of convolution layers. In this model the Convolutional Neural Network consists of 3 hidden layers along with filters. Each image first passes through this series of convolution layers with filters, activation function, pooling and fully connected layers. A

SoftMax function is applied in order to classify the class of the image. The training is done with an epoch count of 500 which obtains us an accuracy close to 100%.

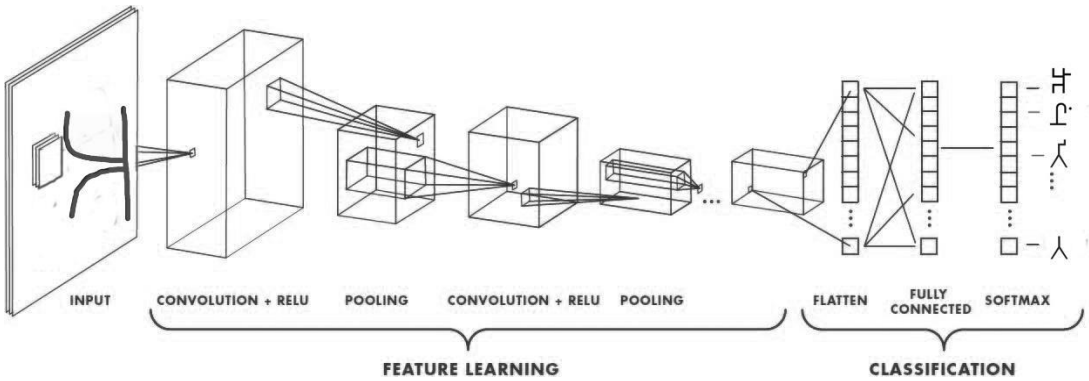

Fig.3 The convolutional neural network of the system.

The Activation function applied in our model is the Rectified Linear Unit. This activation function is preferred as it breaks the linearity in the images by increasing the non-linearity in the neural network. Pooling is done to make sure that the neural network has the Spatial invariance property. In this model max- pooling is done with a stride of 3*3.The trained model is saved as a binary file which can later be used while developing the application.

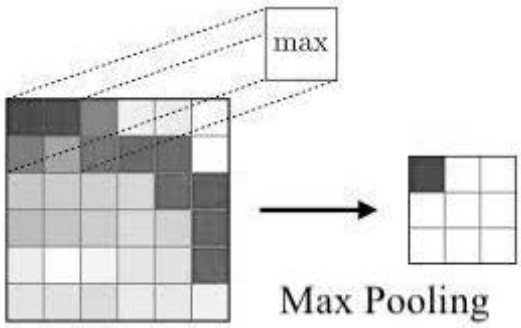

Fig. 4 Max Pooling in Feature Extraction of Image Data

Flask framework is used for developing the application. The image is obtained from the user on the client side and the obtained image is processed using the pre trained Tamil Brahmi character recognition model which classifies the image. The image will fall under any one of the Tamil Brahmi character labels. These labels are then mapped to its equivalent modern Tamil character which provides our desired output.

VII. EXPERIMENTAL RESULTS

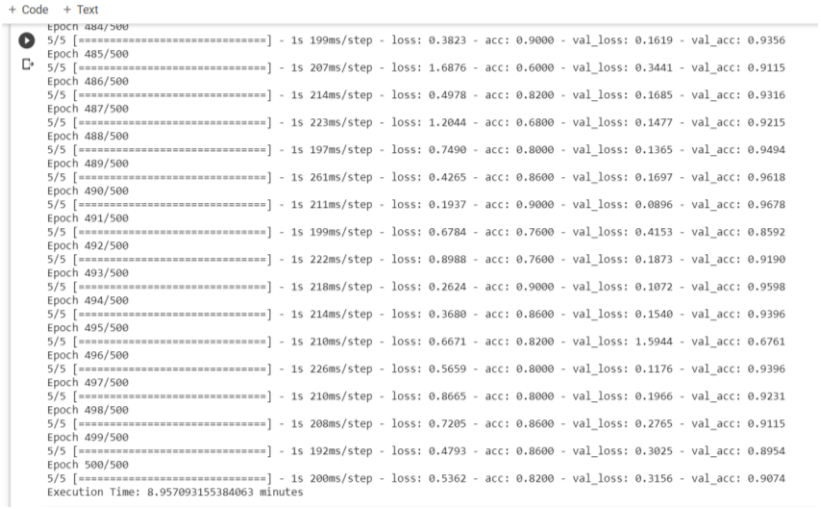

Fig . 5 The TensorLog during training the model showing loss and accuracy.

The prototype of the model gave promising results with an accuracy of 94.6% . The images were fed to the model which was trained and then tested with random inputs. The test results were true thus proving the obtained accuracy.

VIII. FUTURE ENHANCEMENTS

This model is to be trained with the complete, rich, handwritten Tamil-Brahmi character dataset. The model could be used to detect Tamil-Brahmi inscriptions and scriptures which when used with enhanced image recognition algorithms could detect the inscriptions, even from rock engravings and manuscripts.

IX. CONCLUSION

This Character recognition model for interpreting Tamil Brahmi characters aims to be an OCR and hence recognizes the characters from scripts. The input images are processed, extracted and trained to recognize Tamil-Brahmi Characters without errors. This Tamil-Brahmi Character Recognition model has an accuracy of 94.6 %. Further, the model is to be trained with the complete Tamil Brahmi character dataset and is to be experimented with different algorithms to improve the model so that maximum accuracy could be achieved.

REFERENCES

[1]. E.K.Vellingiriraj, M.Balamurugan, and P.Balasubramanie "Tamil Ancient Document Using Machine Translation in Image Zoning".

[2]. T.S.Suganya, S.Murugavalli "An Efficient Ancient Tamil Script Classification System using Gradient Boosted Tree Algorithm".

[3]. N. Sridevi Research Scholar,P. Subashini, Phd. Associate Professor "Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques".

[4]. Neha Gautam and Soo See Chai Faculty of Computer Science and Information Technology, University Malaysia Sarawak "Optical Character Recognition for Brahmi Script Using Geometric Method".

[5]. C.Sureshkumar Department of Information Technology, J.K.K.Nataraja College of Engineering, Namakal,Tamilnadu,India. Dr.T.Ravichandran Department of Computer Science & Engineering, Hindustan Institute of Technology, Coimbatore, Tamilnadu, India "Handwritten Tamil Character Recognition and Conversion using Neural Networks"

[6]. Neha Gautam, R.S. Sharma and Garima Hazrati "Handwriting Recognition of Brahmi Script (an Artefact): Base of PALI Language"

[7]. S. Mageshwaran Department of Computer Science and Engineering, Anna University, G. Alagumalaikannan Department of Computer Science and Engineering, Anna University, India -C. Ravishankar Department of Computer Science and Engineering, Anna University, India M. Kavin Department of Computer Science and Engineering, Anna University, India , S. S. L. DuraiArumugam

Department of Computer Science and Engineering, Anna University, India - "Conversion of Early Tamizh Brahmi Characters into Modern Tamil Characters Using Template Matching Algorithm"

[8]. Ajay P. Singh and Ashwin Kumar Kushwaha Department of Library and Information Science, Banaras,Hindu University, Varanasi – 221005, India "Analysis of Segmentation Methods for Brahmi Script"

[9]. Megha Agarwal, Shalika, Vinam Tomar, Priyanka Gupta "Handwritten Character Recognition using Neural Network and TensorFlow"

[10]. E.K.Vellingiriraj,Dr. P.Balasubramanie Dept. of CSE, Kongu Engineering College, Perundurai, Erode, Tamilnadu, India. "Recognition of Ancient Tamil Handwritten Characters in Historical Documents by Boolean Matrix and BFS Graph"

