International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217230

Recognition of Ancient Tamil Characters from Epigraphical inscriptions using Raspberry Pi based Tesseract OCR

M. Merline Magrina
Assistant Professor, Department of ECE, Er. Perumal Manimekalai College of Engineering, Hosur, TamilNadu, , India

ABSTRACT

Optical Character Recognition (OCR) is the process of identification of the printed text using photoelectric devices and computer software. It converts the inscribed text on the stones into machine encoded format. OCR is widely used in machine learning process like cognitive computing, machine translation, text to speech conversion and text mining.OCR is mainly used in the research fields like Character Recognition, Artificial Intelligence and Computer Vision. In this research, the recognition process is done using OCR, the inscribed character is processed using Raspberry Pi device on which it recognizes characters using Artificial Neural Network. This work mainly focuses on the recognition of ancient Tamil characters inscribed on stones to modern Tamil characters belong to 9th and 12th century characters. The input image is subjected to gray scale conversion process and enhanced using adaptive thresholding process. The output image is subjected to thinning process to reduce the pixel size of the image. Then the characters are classified using Artificial Neural Network Architecture and the classified characters are mapped to modern Tamil character using Unicode. The Artificial Neural Network has input layer, hidden layer of 15 neurons and output layer of 1 neuron to classify the characters. The accuracy of the constructed system for the recognition of epigraphical inscriptions is calculated. The above process is carried out in raspbian environment using python process.

Keywords : ANN, Character Segmentation, Character Recognition, Open CV-python, Raspberry Pi, Tesseract OCR, Unicode values.

Article Info
Volume 7, Issue 2
Page Number: 264-270

Publication Issue :
March-April-2021

Article History
Accepted : 25 March 2021
Published : 31 March 2021

I. INTRODUCTION

Pattern Recognition is the process of recognizing the patterns using Machine Learning algorithm (ML). It is the classification of data based upon the statistical information gained about those patterns. In Pattern Recognition (PR) the input data image is preprocessed, segmented, features are extracted, based upon the extracted features the data is classified using suitable classifier and the respective data is mapped or recognized. The pattern recognition is recognized into 2 important aspects namely classification & clustering (Fig1).

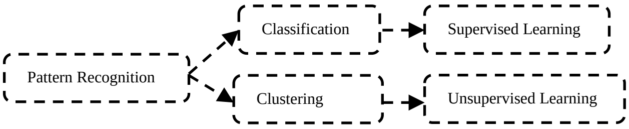


Fig 1 : Aspects of Pattern Recognition

Optical Character Recognition (OCR) is a technology that is widely used for recognizing the text in images i.e. scanned documents, handwritten documents, captured images, printed documents. OCR is used for converting the written text into computer editable form or machine codes. OCR is of two types namely Online & Offline Character Recognition (Fig 2), our proposed work mainly focuses on the recognition of text in Offline format i.e. text from the stone manuscript (Fig 3).

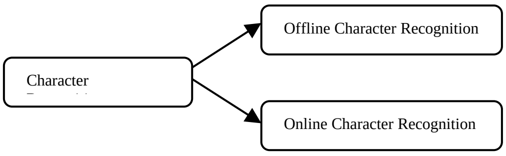


Fig 2: OCR Types

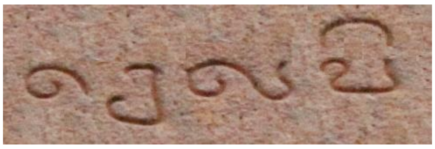


Fig 3: Stone Manuscript

The recognition of Ancient Tamil character from the stone manuscript to the modern Tamil character is performed using OCR methodology. As a part of our work the software development Open CV (Open Source Computer Vision) libraries are utilized to capture the stone manuscript for character recognition. We also try to make it a low cost, light weight mini embedded Raspberry Pi computer for the stone inscription recognition. To provide processed results, this use a simple algorithm & open source OCR called Tesseract OCR on Raspberry Pi[18]. For further processing our dataset images must be divided into Training & Testing set (Fig 4). Generally the 80% of the dataset is treated as a Training set that is used to build a model in order to extract features for training the constructed model (neural network) and the remaining 20% of the dataset is treated as a Testing set is used for testing the trained model in order to calculate the accuracy of the system.

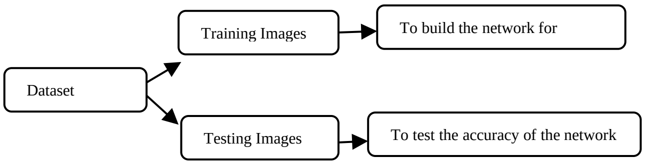


Fig 4: Learning Pattern Recognition

Ancient & Modern Tamil Characters

Testimonials plays important role in historical formation. The historical loyalties [6] are written only through Epigraphical testimonials like stone, temple walls caves, rocks, palm scripts, copper plates, coins, conch, etc. The proposed work concentrates on the stone inscriptions that belong to the Chola Period that is considered as a Golden Period. From this stone inscriptions we came to know about their humanity, dharma, religious bouquets, culture, literature & music. Stone inscriptions are the milestones of Chola period in TamilNadu history. It is not very easy to read the stone inscriptions without the knowledge of linguistics, language, literal & historical backgrounds. This can be read out very easily by practicing these words or characters. The characters found in stone manuscripts is shown in Fig 5,

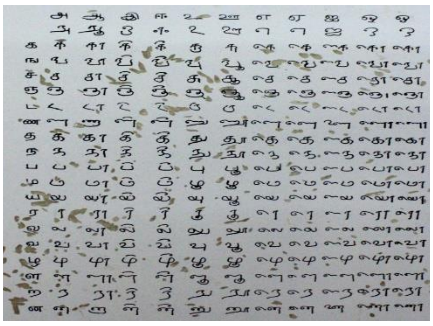


Fig 5: 12th century characters in stone manuscript [6]

Beauty of 12th century Tamil characters [6]:

1. The consonants are bare in nature.
2. The sentences are written continuously without any full stop at the end of the sentences.
3. If the letter comes twice the first letter is considered as pulli consonant.
4. No rules are followed for distinguishing egara, eegara, yegara, yegaara, ogara, oogara letters.

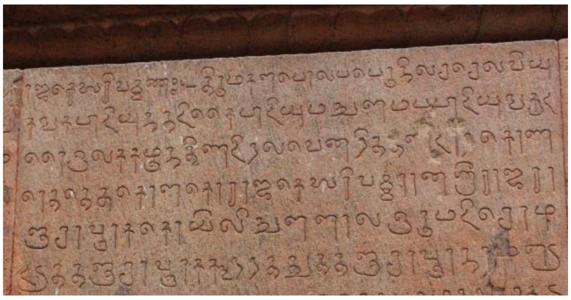


Fig 6: Sample stone manuscript [6]

Explanation of the above manuscript

ராஜ தேவர்க்குயாண்ட திருமகள் போலப் பெருநிலைச்செல்வியு
கங்கபாடியு நீ தடிகை பாடியும் நுளம்பாடியுங் குட
ரை இலக்கழுந் திண்டிரல வெனரித் தளடார கொண்
ரெத் தெசு கோள் கோ ராஜ தேவர்க்குரான ஸ்ரீ ராஜ ரா
ஞ்சாவூர்க் கோயிலிநுள்ளால் இருமடி சோழ
டடுத் தஞ்சாவூர்க் கூற்றத்துத் தஞ்சாவூர் நாம் எடு

Modern Tamil characters are 12 vowels, 18 consonants and 216 composite characters & 1 special character called aayutha ezhuthu counting a total of 247 characters (Fig 7).

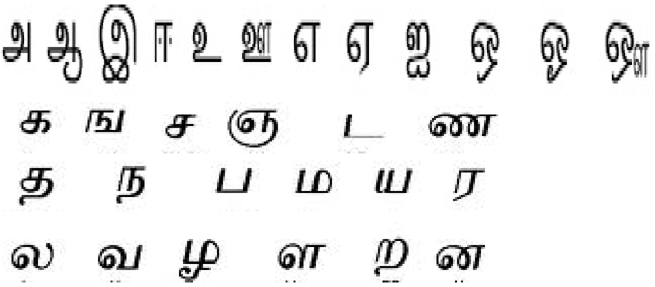


Fig 7: Modern Tamil Script

II. LITERATURE SURVEY

Anush Goel and Akash Sehrawat [1] presents a system for automatic document reader for visually impaired people using Raspberry Pi. It uses OCR technology to identify the printed characters using image sensing device & python programming. The images are converted into audio format using OCR & Text-to-speech synthesis (TTS). The conversion of image into text is done using Raspberry Pi which again uses Tesseract library & python programming. The text files are processed using OpenCV library & audio output is achieved.

Elumalai and Sundar Rajan [2] designed a module that either uses a webcam or mobile camera that is linked with the Raspberry Pi to focus on a range of text. OCR package installed a Raspberry Pi tests into a digital article which is then subjected to skew modification, segmentation & feature extraction. This automatically focuses the region of text in the object, the characters are localized using localization algorithm that uses feature identification & edge pixel distribution using ANN. The text characters are the binarized into a machine editable format using OCR called Tesseract. The recognized characters are then converted to audio format which can be widely useful for the blind persons.

Thiyagarajan and Saravana Kumar [3] developed a system to assist blind persons to read text from the challenging pattern and document. The input image is captured by using a web cam within the Raspberry Pi & follows the OCR steps. The automated system will check the document & read out the substance. The vocal is delivered with the speaker assistance that could help the individual to read the content. This type of reading does not consume more space but results in less recognition rate due to the adjustment in incorrectly spelled words.

Deals with the recognition of English & Chinese characters from images using Corner Detection (Canny Detection) and CNN developed by Beihai Tan, Zipei Zhang [5] Classification by SVM classifier and recognized by CNN method. Character detection is divided into 5 categories i.e. texture based approaches, edge detection, connected component approach, corner detection approaches, and machine learning based approaches. The edge detection method called canny is fast and robust in nature, susceptible to interference, noise sensitive, and detection of fast text detection. The method based on connected component is simple & fast and suitable for Chinese character detection but has poor effect on low contrast character pictures & background. It needs less storage & suitable for hieroglyphics. Machine learning method is time consuming one and poor in versatility. Harris corner detection method is used, the main idea is that if the pixels around the display are more than in one direction of the edge, that point is called corner. It simply explains a matrix similar to autocorrelation function & calculates the 2 eigen values representing curvature value, if they are higher than the threshold value then the point is called corner. Recognition rate obtained is 72% and purely depends upon the character detection results.

T.K. Das and Asis Kumar Tripathy [7] proposed a system for recognizing the characters with high accuracy by means of ANN that involves simple edge detection & matching with the template data. Here the characters are recognized even when noise such as inclination & skewedness presents by training the network to look for discrepancies in data. The proposed work concentrates on NN approach the task of OCR, but it is not effective for feature extraction that is unsuitable for handwritten documents. The feature extraction method can be changed to suit the needs of multiple languages. Such work can be used to read hieroglyphics & other ancient languages.

Analyzing and recognition of ancient Tamil characters from the epigraphical inscriptions and converting them to digital format which utilizes SVM classifier for classification and ANN with Back Propagation Model for recognition by K. DurgaDevi, Dr. UmaMaheswari [8] The research is to develop a technique to capture the inscriptions directly from the stone using digital camera with 3D resolution, which is very easy & fastest approach, developed mainly for preserving in digital format used efficiently by epigraphers. HMM model is well suited for providing good accuracy rate for Chinese character recognition by considering geometric feature extraction. Several methods have been involved in this system for feature extraction & classification of recognized pattern for national languages i.e. Hoyasala, Marathi, Devanagiri & Tamil. Since Tamil languages not involved with much curliness, generally the structural & statistical features are more concentrated. SVM & ANN with BP model and KNN classification model is widely applied to all scripts. It has more computational power, convergence & capacity. The accuracy of recognition is about 66%. This recognition of characters purely depends upon input image and detection method.

Recognition process is done using OCR the character code in the text are pre-processed using Raspberry Pi device on which it recognizes the character using Tesseract algorithm & python programming by Mallapa P. Gurav and S. Salimath [9]. OCR performs Document Image Analysis (DIA) for virtual digital library design & construction. Raspberry Pi features a Broadcom system on chip (SOC) which includes ARM processor compactible with CPU & GPU. Using Tessaract library the image will be converted into data & detected data will be displayed on the monitor and also it will be pronounced through earphones using Flite library.

Praveen Kumar and Padmaja [10] proposed a character recognition system for English language in Online OCR model & the designed process is implemented with the Raspberry Pi chip. The character is recognized using Fast Artificial Neural Network (FANN) and the Linux OS environment is utilized. This type of OCR requires the prior knowledge about the characters and need more training on the character samples. The recognition rate of 73% is obtained using normal Raspberry Pi board. The advanced Raspberry Pi board will help the growth of the project & successfully implemented for testing more number of English alphabet samples.

Rajasekar. M and Celine Kavitha proposed a method [11] to recognize the handwritten Tamil characters using ANN. This describes the behaviour of different models of Neural Network used in OCR. In this paper the parameters like Hidden Layer, size of Hidden Layer & epoch are found. In pre-processing they have applied some basic algorithms for segmentation of characters, normalizing the characters & De-skewing. OCR is aimed at recognizing printed document. The input document is read pre-processed, feature extracted and recognized text is displayed in a picture box. OCR eliminates the difficulty by making the data available in the printed format. After implementing an OCR for Tamil character text for few words, used to check the whole paragraph for implementation. Then this is applied for different dataset for the testing process and the accuracy of the network is obtained.

Sushama Shelke and Shaila Apte [12] developed a real time recognition system for character recognition of Indian Scripts of Marathi language. The developed system uses Raspberry Pi & OpenCV python programming. The system comprises of 2 modules namely image acquisition & recognition. The image acquisition consists of pre-processing steps and the recognized characters are converted into speech. The recognition rate of 92 % is obtained. The template matching fails if the original image area is smaller than the template area. Template matching fails in case of translation & rotation.

A system has been proposed to recognize the ‘hand’ written Tamil characters from Historical Documents using ‘Boolean Matrix’ by Vellingiriraj & Balasubramani [13] This is effectively used for recognizing the characters from the stone images. The self adaptive system uses features which enrich the system accuracy. Finally the proposed method handles segmentation, smoothing, filtering, and normalization of complex ancient Tamil character recognition from stone inscriptions.

Offline character recognition method [14] proposed by Dr. Gunasekaran & Preethi the pre-processing technique results better accuracy such as noise removal, gray scale conversion & binarization are employed to enhance the character before classification. Segmentation is used to split the characters to single part and features are extracted with transfer edge technique. Neural Network is implemented to recognize the character. This can be improved with different classifiers such as SVM, SOM, Tree classifier, Fuzzy classifier and compares the accuracy.

Ajantha Devi and Santhosh Baboo [15] proposed a system of OCR for visually impaired people who cannot read the text document but need to access the content of the document. This utilizes the camera based assistive device that can be used by people to read the Tamil text which is implemented by capturing the image in an embedded system based on Raspberry Pi to recognize the Modern Tamil script. The synthesized character is converted into audio format & played using microphone connecting to on-board audio jack of Raspberry Pi.

Pranjali Pohankar and Namrata Taralkar [16] proposed a system for recognizing the characters written on paper documents converting it into digital form. Neural Network is a technique used to improve the accuracy & efficiency of the Handwritten Character Recognition. The error back propagation algorithm is used to train the MLP network. The advantage of BPN is that it can fairly approximate a large class of functions. The variations in the handwritten character reduce the recognition rate and mainly require the filtering and feature extraction methods to be accurate.

Ankit Sharma and Dipti R. Chaudhary [17] proposed a system to recognize the Offline Handwritten English Characters using Feed Forward Back Propagation Neural Network. The ANN is trained using the Back Propagation algorithm to recognize the English Numerical numbers represented in binary form. The system has a degraded performance characteristics in feature extraction that lead to higher error rates in classification methodology. From this the noise removal process is an essential one in the pre-processing step of OCR.

III. METHODOLOGY

The proposed work is to recognize the 12th century characters of the stone manuscript using Raspberry Pi camera module, the captured image is applied to the Tesseract OCR [18] [19] of Artificial Neural Network (ANN) module and the classified characters are matched with the Modern Tamil character using Unicode. The proposed block diagram of our work is shown in Fig 8,

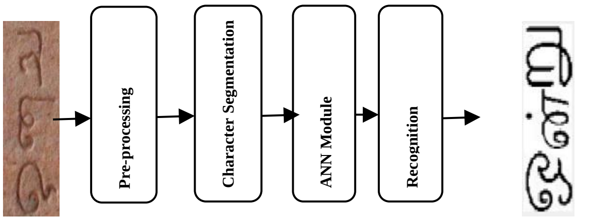


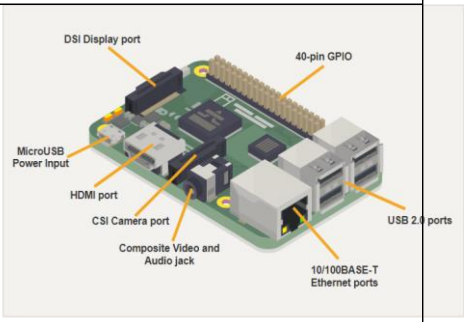
Fig 8 : Proposed block diagram

<u>Raspberry Pi 3 B+ Module</u>:

The Raspberry Pi 3 Model B+ (Fig 9) is a tiny credit card size computer. Just add a keyboard, mouse, display, power supply, micro SD card with installed Linux Distribution and you'll have a fully fledged computer that can run applications from word processors and spreadsheets to games. The Raspberry Pi 3 Model B+ is the first Raspberry Pi to be open-source from the get-go, expect it to be the de-facto embedded Linux board in all the forums.

Fig 9: Raspberry Pi 3B+ module

Specification of the Raspberry Pi [20]

Chip	Broadcom BCM2837 64 bit ARMv8 Quad core Cortex A53 1.2 GHz
Storage	microSD Card
Memory	1 GB
Graphics	400 MHz Dual Core Videocore IV GPU OpenGL ES 2.0 Hardware Accelerated OpenVG 1080p30 H.264 high-profile decode
Weight	42g
Audio	HDMI port supports multichannel audio o/p Audio line out/3.5mm headphone jack
Connections	 DSI Display port 40-pin GPIO MicroUSB Power Input HDMI port CSI Camera port Composite Video and Audio jack USB 2.0 ports 10/100BASE-T Ethernet ports
Communication	802.11n – Wi-Fi wireless networking Bluetooth 4.1 wireless technology 10/100BASE-T Ethernet
Electrical & Operating Requirements	Input voltage: 5V DC Current Requirement: 2.5A
Operating Systems	NOOBS & RASBIAN

Raspberry Pi Camera Module:

The camera module used in this project is Raspberry PI CAMERA BOARD as shown in the Fig 10. The camera plugs directly into the Camera Serial Interface (CSI) connector on the Raspberry Pi 3 B + module. It's able to deliver clear 5MP resolution image, or 1080p HD video recording at 10fps.

Fig 10: Experimental Setup

Artificial Neural Network Architecture: [19]

Neural Networks are the building blocks of Artificial Systems that are inspired by biological neural networks. They are based upon the threshold logic computational model. They also based either on the study of brain or application of Neural Networks to Artificial Intelligence [4]. The main components of ANN are neurons, connections, weights, biases, propagation functions & learning rule. This neural network may be either supervised or unsupervised. Supervised Machine Learning has an input neuron variable x and output neuron variable y. The applied algorithm learns from the training dataset [20]. With each correct output the algorithm iteratively makes the predictions. The learning of the network stops when it reaches the desired level of performance. This is for classification & regression. Whereas in Unsupervised Machine Learning has an input neuron x and no output neuron y. This is to model an underlying structure of the data for studying the structure when it is training. This is for clustering & association. The steps followed in training the network is take the inputs (I_i), multiply it with weights (W_i),

$Y = W_i * I_i$

The calculated results (Y) are passed through a Sigmoid functions to calculate the output, this function is to normalize the results in the range if 0 and 1.

$R= 1/ 1+ e^{-y}$

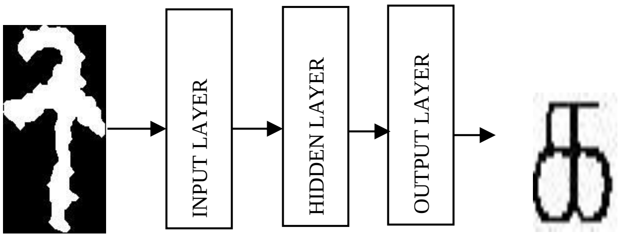


Fig 11: Artificial Neural Network Architecture

.Results and Discussion

The captured input image of dimension 357 * 133 is fed to the Tesseract OCR module that is resized to fixed size of 350 * 800 dimensions as shown in Fig 12

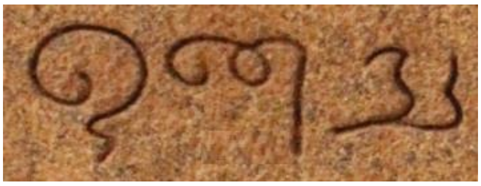


Fig 12: Input image

The RGB 3 plane image is converted to 2 plane gray scale image that retains luminance component in order to increase the rate of accuracy as shown in Fig 13

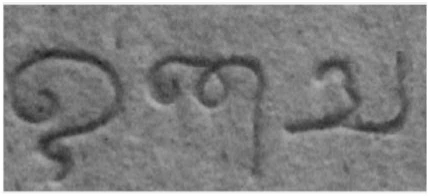


Fig 13: Gray scale image

The gray scale image is subjected to enhancement and binarized to remove noises in the image as shown in Fig 14

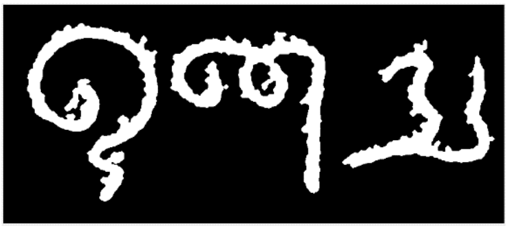


Fig 14: Noise Removed image

From the noise free image the pixel size of the characters are reduced by the process called thinning as shown in Fig 15

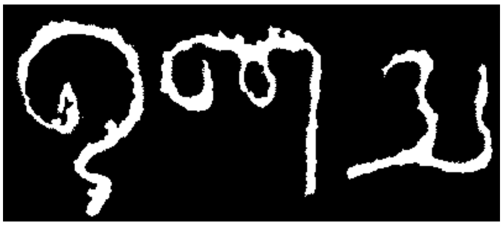


Fig 15: Thinning Image

The characters are classified by the trained ANN architecture and the classified characters are matched/mapped using Unicode of the exact Modern Tamil characters [50, 19, 20] as shown in Fig 16

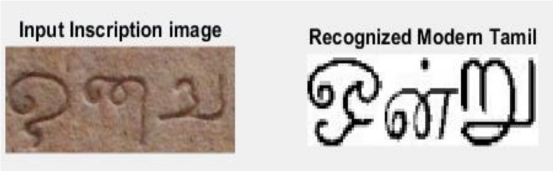


Fig 16: Recognized Modern Tamil Character

Table 1: Performance metric of the system

Metric	**Measure**
Recognition rate (RR)	99%

IV. CONCLUSION

Thus the proposed method focuses on the recognition of ancient Tamil characters from the epigraphical inscriptions based on Tesseract OCR using Raspberry Pi. The input image is processed through gray scale conversion, adaptive thresholding, thinning process to remove noises from the image. The ANN architecture is implemented to classify the characters and the characters are mapped using the Unicode values. The Recognition rate 99 % is obtained using Raspberry pi is and it is found to higher than the other models.

V. ACKNOWLEDGMENT

I owe my sincere thanks to **Dr. M. BHAVANI, Assistant Professor**, Department of Epigraphy & Archaeology, Tamil University, Thanjavur and **Mr. M. RAJENDRAN, Record clerk**, Department of Archaeology, Thanjavur who guided me and encouraged me to do the project in each and every step.

The author thanks the Government of TamilNadu for having selected the project under the project code EEE-013 of TamilNadu State Council of Science & Technology (TNSCST 2018-2019). I am immensely thankful to the TNSCSCT for having furnished me with the financial support in order to complete the project work .

VI. REFERENCES

[1] Anush Goel, Akash Sehrawat, Ankush Patil, Prashant Chougule and Supriya Khatavkar (2018), "Raspberry Pi based reader for blind people", International Research Journal of Engineering and Technology (IRJET 2018), vol: 5, issue: 6, pp: 1639- 1642.

[2] Elumalai .G, J. Sundar Rajan, P. Surya Prakash, V.L. Susruth and P.K. Sudharsanan (2018), "Design and Development of Tessaract– OCR based assistive system to convert captured text into voice output", International Research Journal of Engineering and Technology (IRJET 2018), vol: 5, issue: 4, pp: 509 – 513.

[3] Thiyagarajan, Saravanan Kumar, Praveen Kumar and Sakana (2018), "Implementation of Optical Character Recognition using Raspberry Pi for visually Challenged People", International Journal of Engineering & Technology (IJET 2018), vol: 7, issue: 3, pp: 65 – 67.

[4] Vishwanath Bharadwaja, Ananmy, Sarraf Nikhil, Vineetha (2018), "Implementation of Artificial Neural Network on Raspberry Pi for signal processing applications", International Conference on Advances in Computing, Communications & Informatics (ICACCI 2018), pp: 1488 – 1491.

[5] Beihai Tan, Chao Hu and Zepei Zhang (2017), "Character Recognition based on Corner Detection", IEEE International Conference on Natural computation, Fuzzy System and Knowledge Discovery (ICNC-FSKD), pg: 503-507.

[6] Bhavani. B (2017), "Tamilnadu Historical Documents (Epigraphical Inscriptions)".

[7] Das T.K., Asis Kumar Tripathy and Alekha Kumar Mishra (2017), "Optical Character Recognition using Artificial Neural Network", International Conference on Computer Communication & Informatics (ICCCI 2017), pp: 1-4.

[8] DurgaDevi. K and Uma Maheswari (2017), "Insight on Character Recognition for calligraphy digitization", IEEE International Innovations in ICT for Agriculture and Rural Development (TIAR), pp: 78-89.

[9] Mallapa D.Gurav, Shruthi S.Salimath, Shruthi B.Hatti and Vijayalakshmi I.Byakod (2017), " B-Light A Reading aid for the Blind people using OCR & OpenCV", International Journal of Scientific Research Engineering & Technology (IJSRET 2017), vol: 6, issue: 5, pp: 546 – 548.

[10] Praveen Kumar, Padmaja, Nagadeepa and Sagar Reddy (2017), "Online English Character Recognition using Raspberry Pi", International Journal of Scientific Engineering & Technology

Research (IJSETR 2017), vol: 6, issue: 20, pp: 3978-3981.

[11] Rajasekar. M, Celine Kavitha and Anto Bennet.M (2016), “Performance and analysis of Handwritten Tamil Character Recognition using Artificial Neural Network”, International Journal of Recent Scientific Research (IJRST), vol: 7, pp: 8611-8615.

[12] Sushama Shelke, Shaila Apte (2016), “ Real time character reading system for Marathi script using Raspberry Pi”, 3rd International Conference on Electrical, Electronics, Engineering, Trends, Communication, Optimization & Sciences (EEECOS 2016), pp: 1 – 5.

[13] Balasubramanie.P and Vellingiriraj. E.K (2015), “Recognition of ancient Tamil Historical documents by Boolean Matrix & BFS graph”, International Journal of Computer Science & Technology, vol: 5, pp:65-68.

[14] Preethi. N and Gunasekaran. T (2015), “Language Specific Offline character recognition using Neural Network Classifier”, International Journal of Advanced Research in Electronics & Communication Engineering (IJARECE), vol: 4, pp: 218-221.

[15] Ajantha Devi and Santhosh Baboo (2014), “Embedded Optical Character Recognition on Tamil Text Image using Raspberry Pi”, International Journal of Computer Science Trends & Technology (IJCST 2014), vol: 2, issue: 4, pp: 127 – 131.

[16] Pranjali Pohankar, Namrata Taralkar, Snehalata Karmare and Smita Kulkarni (2014), “Character Recognition using Artificial Neural Network”, International Journal of Electronics Engineering & Computer Engineering (IJECCE 2014), vol: 5, issue: 4, pp: 245 – 248.

[17] Ankit Sharma, Dipti. R. Chaudhary (2013), “Character Recognition using Neural Network”, International Journal of Engineering Trends & Technology (IJETT 2013), vol: 4, issue: 4, pp: 662 – 667.

[18] https://www.pyimagesearch.com/2018/09/17/opencv-ocr-and-text-recognition-with-tesseract/

[19] https://www.geeksforgeeks.org/implementing-ann-training-process-in-python/

[20] https://towardsdatascience.com/how-to-build-your-own-neural-network-from-scratch-in-python-68998a08e4f6

[21] https://www.python-course.eu/neural_networks_with_python_numpy.php

[22] https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/

Cite this article as :

M. Merline Magrina, "Recognition of Ancient Tamil Characters from Epigraphical inscriptions using Raspberry Pi based Tesseract OCR", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 118-127, March-April 2021. Available at doi : https://doi.org/10.32628/CSEIT217230
Journal URL : https://ijsrcseit.com/CSEIT217230