How to Cite:
Harini, M. L., Pandian, A., Hithayathullah, A., & Sadiq, A. K. (2022). Handwritten Tamil character recognition. *International Journal of Health Sciences*, *6*(S5), 2127–2133.
https://doi.org/10.53730/ijhs.v6nS5.9102

Handwritten Tamil character recognition

Harini M L
***Pandian A**
Hithayathullah A
Dr. A.K. Sadiq

***Abstract*---**This paper focuses on recognizing Tamil characters that are handwritten and displaying their digital variant. Tamil is a regional language and one of the classical languages of India. Tamil literature has a significance of over four thousands years and many of those literatures are yet to be digitized. The main objective of this paper is to convert old handwritten Tamil documents or scriptures into their digital form so that they can be preserved over a long period of time. Character recognition generally can be implemented by using either CNN or OCR methodologies. In this paper we will be using Optical Character Recognition commonly known as OCR for implementation. It is a technique which involves four modules in order to recognize the characters.

***Keywords*---**OCR, CNN, Tamil.

1. Introduction

Tamil is one of the longest surviving languages in the world, oldest of the Dravidian languages and hence considered as one of the classical languages of India. It is considered as an ancient language which is proven to have a lifespan of over four thousand years. The significance of Tamil language is that it is the only language that has history for more than two thousand years with a rich literary tradition . It is being spoken in the southern parts of India and Sri Lanka. Tamil language has been declared as the official language of the state Tamil Nadu and the union territory Pondicherry. There is a significant amount of Tamil speakers in Malaysia and Singapore.

Tamil literature is the most important, with a history spanning over four thousand years. Between the 3rd century BCE and the 2nd century CE, Tamil literature was classified as Sangam literature, and this period is known as the Sangam period. The Sangam period is known as the Golden Age of Tamil Literature because it plays such a vital role in Tamil culture. Silappadhigaram, Tholkaappiyam, Manimegalai, Valayaapathi, and Seevaga-Sinthaamani are the five epic literatures of Tamil. More Tamil literatures that are centuries old have

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 18 March 2022, Manuscript revised: 9 April 2022, Accepted for publication: 27 May 2022

yet to be digitised. We will use OCR technology to convert handwritten Tamil papers into digital form in this article. The Tamil alphabet is divided of 12 vowels, 18 consonants, 216 composite characters, and one unusual character (ak), which will be utilised as the dataset.

Character recognition is one of the most difficult disciplines of research since the efforts made so far to develop it have not totally solved the majority of the challenges that are both intellectually and commercially important. The process of transferring a language's characters from their spatial representation to a digital symbolic representation is known as handwritten character recognition. Researchers have been working on Handwritten Character Recognition for the past thirty years. Given how difficult it is for humans to read handwritten texts with 100 percent accuracy, the idea of developing a totally accurate Handwritten Character Recognition System appears impossible.

2. Handwritten Tamil Character Recognition

The method of detecting scanned handwritten characters and showing their digital equivalents is known as handwritten character recognition. Online and Offline Character Recognition are the two main types of Handwritten Character Recognition. The automatic conversion of characters written on a tablet, a dedicated digitizer, a PC, or a PDA is referred to as online handwritten character recognition (Personal Digital Assistant). To distinguish handwritten notes, a sensor is utilised to capture pen tip movements as well as pen switching.

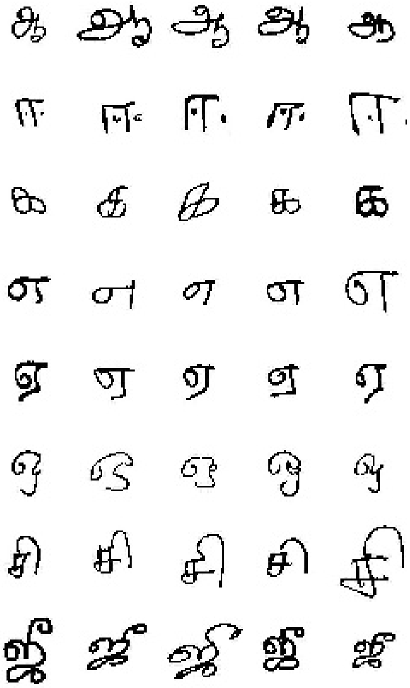


Fig.1 Scanned copy of handwritten Tamil characters

Offline Character Recognition is a process that deals with obtaining a scanned handwritten document which is used as the dataset for conversion. In this paper, we train and identify offline Tamil handwritten character datasets using a Convolutional Neural Network (CNN) based Optical Character Recognition (OCR). After scanning and saving a handwritten Tamil document in TIFF, JPG, or GIF format, pixel processing or pre-processing is conducted. Pixel processing is carried out in three stages. Binarization which is the transformation of grey scale images into black and white image samples. Noise reduction is the technique of filtering image samples to remove noise that happens during transmission, photocopying, and image degradation.. Segmentation is performed on the scanned document after which it undergoes feature extraction. Scanned handwritten samples are provided as an input to a dynamic programming algorithm, which recognizes and matches them with their respective digital character datasets. As the main aim of this system is to preserve old Tamil literature and documents, we take Tamil literature and old Tamil poems into consideration for our datasets.

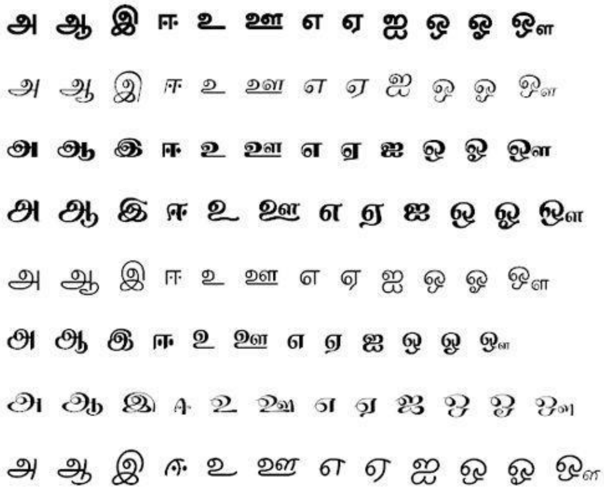


Fig.2 Digital Tamil characters

Fig.1 shows the scanned copy of handwritten Tamil characters and Fig.2 shows their digital equivalent. The major challenge here is that Tamil is a regional language which has a huge variation of characters and each individual has unique writing styles. This makes it difficult to recognize certain characters. Increases in the scale of mis-recognition and errors were experienced. Hence we work on training separate dataset models for each individual's handwriting. This way we ensure a low mis-recognition rate and more accurate character recognition. Each word that has been trained is called a corpus and when an image is provided as an input the OCR compares the image with the trained corpus and based on the best matching rate, it displays the digitized copy of the input provided.

3. Modules

In this paper we use four modules for recognizing handwritten characters and displaying their digital form as output. Scanning, Pre-processing or pixel processing, Image Segmentation, Feature extraction and Recognition are the main modules used. These modules undergo certain processes within each in order to arrive at the final result. The architecture diagram which explains the working pipeline of our system is shown in Fig.3

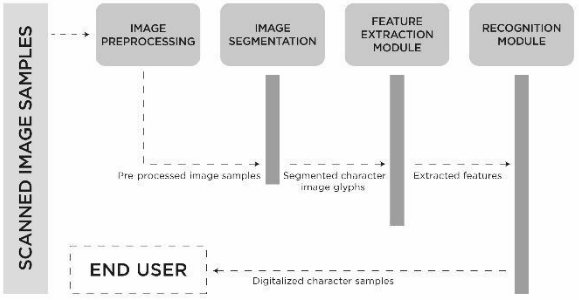

SCANNED IMAGE SAMPLES
IMAGE PREPROCESSING
IMAGE SEGMENTATION
FEATURE EXTRACTION MODULE
RECOGNITION MODULE
Pre processed image samples
Segmented character image glyphs
Extracted features
END USER
Digitalized character samples

Fig.3 Model architecture Diagram

In the proposed system, we take scanned image samples of handwritten Tamil scripts. Here we use our own handwritten copy of a single individual for training the dataset models. The scanned image is preprocessed where it undergoes binarization, noise removal and skew correction in order to convert the input image into grey image and make it suitable for further processes. The resultant image is segmented into paragraphs, paragraphs into lines, lines into words and words into character image glyph. The segmented image samples are subjected to a normalization process, which converts the variable sized character image glyphs into standard size. The features such as character height, width, number of horizontal and vertical lines, horizontally and vertically oriented curves, number of circles, number of slope lines, image centroid, and special dots are extracted from each character image glyph. After the features of the image have been extracted each character is assigned to a unique vector which acts as an identity for the particular character. It is then sent to a recognition module which compares the input sample with the available number of trained dataset corpuses and based on the best matching score it prints the digitized variant of the handwritten input sample as an output to the end user. The basic working of the OCR has been explained. Each module mentioned above undergoes several sub modules or sub processes which are detailed below.

3.1 Scanning

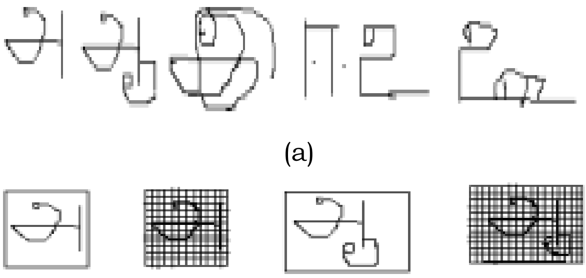
Scanning is the initial phase, which involves digitising a real-world image. A handwritten paper is scanned, and the image is subjected to random pixel value changes throughout each step of the scanning process, which is known as noise. The scanned document is then saved in one of three formats: TIFF, JPG, or GIF.

3.2 Pre-Processing

The process of processing a scanned image for the first time is known as image pre-processing. Binarization, Noise removal, and Skew correction are the three primary processes in this procedure. Binarization is the process of converting an image into a greyscale image (black and white). Here, the image is brightened first, then binarized. There are two peak values in the grey image samples: a high peak and a low peak value. The high peak value corresponds to the grey image's white backdrop, while the low peak value refers to the grey image's foreground. The pre-processed binarized image sample is used to remove noise. Noise can be created by a variety of sources, including poor scanned document quality, accumulated white scanning, transmission, photocopying of photos, and so on. The noise in the image must be removed before it can be further processed, which is done by filtering the image. After noise removal, the image is tested for skewing. Skew correction is the process of checking and correcting the image's orientation. A picture can be skewed either to the left or to the right. During skew correction, the image is tested for a 15-degree angle of disorientation, following which the image is rotated till the lines align with their actual horizontal axis.

3.3 Segmentation

After pre-processing, the noise free and skew corrected image undergoes segmentation where the image contents are being split up into individual characters.



(a)

(b)

Fig. 2 Segmentation (a) Original image (b) Segmented image

Segmentation is performed majorly to group pixels into image regions in order to extract proper samples for further analysis. Segmentation not done right can result in mis-recognition or rejection of salient image regions. In this phase the binarized image is checked for inter line spaces, which if detected can be used as a guide for segmenting the image into paragraphs. The lines in the paragraphs are then checked for horizontal space intersection with the background in order to segment them further. Segmentation between the lines in a paragraph can be determined by scanning from the first row of the document. If there is a difference in the black pixel number between two rows, it is indicated as a new line of text. The width of the horizontal space intersection can be detected by using the histogram of the image. After segmentation of lines from the paragraphs, these lines are further segmented into separate words by scanning them for vertical space intersection. In order to segment words from the lines we use page layout and character separation. Segmentation of sub-words can be carried out similarly. The decomposition of these segmented words into characters is done via character width computation. Fig.2 displays the original image and image after segmentation.

The resultant character image is scaled so that it could fit a 64x64 window after which it is subjected to a thinning algorithm. After which an algorithm is performed in order to produce a thinned image sample which can further be used for feature extraction. The characters before and after applying the thinning algorithm is shown in Fig.3

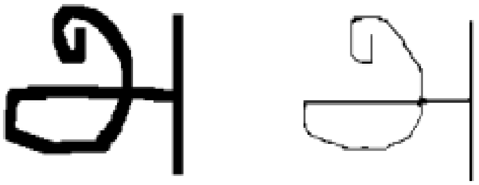


Fig.3 Before and after samples of thinning

3.4 Feature Extraction

The recognition phase's backbone is feature extraction. Feature extraction's major purpose is to extract features from characters such as the character's height and width, the number of horizontal and vertical lines contained in the character, dots, curves, arcs, vertical and horizontal orientation of the characters, and so on. Other factors, such as the image's centroid and pixels in various regions, can aid in increasing the recognition rate. During this phase, each character is represented by a feature vector that serves as the character's identity, making recognition easier.

Complete pipeline of OCR

The scanned image of a Tamil literature document is passed as an input. This image is being pre-processed and passed on to the segmentation module. The segmented image undergoes feature extraction and the resultant image is classified using the proposed convent. Each scanned document has their own individual trained datasets and finally these trained corpuses are used for generating the digitized form of the input provided.

5. Conclusion

In this paper we conclude that OCR technology can be used for digitizing and preserving old and brittle Tamil literature and documents because of their robustness to font size and style, image quality, contrast etc. We have used the existing OCR techniques to convert handwritten Tamil documents into digital form. We have arrived at the goal of digitizing handwritten Tamil scripts by segmenting the script into separate words and then decomposing the individual characters for the feature extraction module. To improve further, research can be done on letter segmentation and identification. In order to achieve this, we consider using a selective search algorithm followed by CNN.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Harini M L and Hithayathullah A; methodology, Harini M L; software, Hithayathullah A and Harini M L; formal analysis, Hithayathullah A; resources, Harini M L; data curation, Hithayathullah A; writing—original draft preparation, Harini M L; writing—review and editing, Hithayathullah A; visualization, Harini M L; supervision, Dr.Pandian A.

References

[1] Tamil Handwritten Character Recognition Using Artificial Neural Network, International journal of scientific & technology research volume 8, issue 12, Ms.G.Thilagavathi, Ms.G.Lavanya, Dr.N.K.Karthikeyan, 2019.

[2] Handwritten Tamil character recognition and Digitalization using deep learning, International Journal of Advanced Science and Technology Vol. 29, No. 3s, N. Sasipriyaa, K.Abirami, G.Banupriya, S.Dhivya, 2020.

[3] offline Handwritten Tamil Benchmarking on Character Recognition using convolutional neural networks, Journal of King Saud University - Computer and Information Sciences**,** Kavitha B R, Srimathi C, 2019.

[4] Handwritten Tamil Character Recognition Using ResNet, International Journal of Research in Engineering, Science and Management Volume-3, Issue-3, R. Jayakanthan, A. Hiran Kumar, N. Sankarram , B. S. Charulatha, Ashwin Ramesh, 2020.

[5] Junction Point Elimination based Tamil Handwritten Character Recognition: An Experimental Analysis, Journal of Systems Science and SystemEngineering volume 29, pages 100–123, M. Antony Robert Raj & S. Abirami, 2019.

[6] Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. "Improving neural networks by preventing co adaptation of feature detectors." arXiv preprint arXiv:1207.0580 (2012).

[7] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012.

[8] LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. "Backpropagation applied to handwritten zip code recognition." Neural computation 1, no. 4 (1989): 541-551.

[9] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted Boltzmann machines." In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814. 2010..

[10] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.

[11] Suryasa, I.W., Sudipa, I.N., Puspani, I.A.M., Netra, I.M. (2019). Translation procedure of happy emotion of english into indonesian in kṛṣṇa text. *Journal of Language Teaching and Research, 10*(4), 738–746

[12] Suryasa, W. (2019). Historical Religion Dynamics: Phenomenon in Bali Island. *Journal of Advanced Research in Dynamical and Control Systems*, *11*(6), 1679-1685.