Offline Tamil Language Character Recognition for Digitalizing Process

K.Shanmugam, Dr.B.Vanathi

Assistant Professor, Professor & Head

Deaprtment of CSE

SRM-Valliammai Engineering College

Shanmugamk.cse@valliammi.co.in, mbvanathi@gmail.com

ABSTRACT:

Tamil is an Ancient and beautiful language and holds the prestigious stand of classical language declared by UNESCO. Integrating it with the technology is a try to enhance its beauty. Language integrating with technology has always proved to be a boon to the learners. There has been number of research done in integrating English language with technology and this has proved to be wider scope of research in the field of English language teaching itself. Tamil being the fifth most spoken language in India, holding the status of official language in Singapore and Sri Lanka apart from the recognition as minority language in South Africa Malaysia and Mauritius, deserves to be explored, enhanced and simplified for its learners at the very best. This Catered researchers interest towards integrating technology and Tamil language. Current scenario states the research study, of handwritten recognition of Tamil script is available only online. The purpose of this project is expanding its scope to make it available offline as well. Neural network algorithm which can be used to classify and recognize the pattern of language stands as a base to run this project. A Set of Sample Handwritten Tamil Characters are taken as input in the image format, to process the character, and train the Neural Network Algorithm, accordingly to recognize the pattern and convert recognized characters to a Printed document. ASCII character mapping is done to convert handwritten script to computerized text format. Neural Networks are particularly useful for solving problems that cannot be expressed as a series of steps, such as Recognizingpatterns, Classifying them into groups, Series prediction and Data mining. The Segmentation involves spacing method and Feature Extraction is done using zoning method. The Neural Network then attempts to

determine if the input data matches a pattern that the Neural Network has memorized. A Neural Network trained for character recognition is designed to take Tamil character input samples and categorize them into groups. This project concerns with improving the accuracy for composite letters. The main aspect of the project is that researchers have worked on the converting the handwritten text to a computerized text format.

***Keywords:** Neural network, Tamil Handwritten Recognition, Feature Extraction, Segmentation*

1. INTRODUCTION:

Technology has always played a vital role in enhancing various arenas. Integration of technology has always proved to be boon in every aspect of studies. The scope of Tamil language research studies enhances with the integration of technology. Being one of the oldest classical language Tamil deserves to be simplified and reach to everyone who desire to learn it and use it. Tamil language has the pride status of 15th most spoken language in the world. Technology integration in Tamil language facilitates researchers and users to explore the language further.

A physical procedure used to adapt an image signal into an physical image is called image processing[1]. There are two types image processing digital and analog. The most common sort of image processing is photography in which an image is captured or scanned with the use of a camera to form a digital or analog image. In process of physical image creation an image is processed by using the suitable technology based on the input source type wherein in a digital image an image is stored in a computer as a file. The file in the computer is rendered through photographic software to give an output an actual image. The details of a photo like colours shading etc. are all captured at the time of taking a photograph and with the help of a software all these details are transformed into an image.

As the study suggest digital image processing surely takes upper hand comparative to analog image processing. Digital image processing has many advantages over analog image processing computer has perform image processing in a digital method that is what comes as an output in digital image processing. Digital image processing comparatively allows wider range of algorithms for input data at the same time minimizes the problem like build up of noise and signal distortion during processing. There are 6 basic steps involved in fundamentals in digital image processing [2].The various basic steps are as followsImage Acquisition,Image Pre-processing

,Image Segmentation,Image Representation and Description,Image Recognition and Interpretation and Knowledge Base.

2. Scope:

The project is to facilitate officials of Tamil Nadu government people who work on documentation in Tamil Language. It is used in Tamil Nadu and in the places where Tamil is used as official language. This project will also help teacher and learner who would use computerized Tamil text for teaching and learning process. The project involves handwritten Tamil Script and formation of its computerized digital Text. Tamil character handwritten recognition process is used for Digitizing the palm leaf manuscript, historical documents by converting and storingit in a digital format.

3. Related work:

Literature survey is the most important part to proceed with the project before getting started with the project it is necessary to comprehend the factors involved in the project. It is important to know the time factor, economic factors and the strength of the project to understand the viability involved in the project it is important to conduct a literature survey. Literature survey paves way to understand views of other researches regarding the topic related to project and the current development in the field of the project. Literature survey provides the guidance and support the researcher needs. Source through which literature survey are obtained could be journals books or websites. Any information regarding the proposed system is taken into consideration for building the project.

(i).K.Punitharaja and P.Elangopaper[3] is a comprehensive study of Tamil character recognition system development. Attributes which evaluates techniques which undergoes in in Tamil character recognition. This paper Debates about identifying and solving problems that are faced during developing a practical TCR system. Paper shows set off criteria for categorising the TCR techniques and Research based on detailed list of features definitions and extractions with classification methods that are frequently experimented by TCR resources.

Inference:

The paper presents itself with the tabulated recognition result summary that is used to compare and contrast the performance of respective recognition techniques. The result tabulation presented is problem where it doesn't provide a clear view of the data sets used and the state of the system where recognition experiment was conducted. It explains the characteristics of the Tamil language and its character specifications which is suitable for recognition perspective. This analysis is helpful in determining the difficulties that could prove to be a challenge for the handwritten Tamil script recognition system. The unavailability of the test data causes the analysis of comparing and contrasting the techniques used in TCR complex. the factors that affects text production process are: font type, paper texture, paper colour, ink colour and font size.

(ii).D.Rajalakshmi and S.K.Jayanthi paper[4] analyses the use of writer identification by using the textual and structural features of the Tamil language and it has particularly proposed a method called connected component labelling which is used for extracting the character level feature of the text. In this approach segmentation is done for an image of hand-written document into individual characters then it's used for writer identification.

Inference of this paper:

This paper presented a new approach called connected component labelling which might produce inaccurate result if the objects of test is not aligned chance. It might cause computational complexity for objects having many parameters.

(iii).L.Suryakala and Dr.P.Thangarajpaper[5] describes about using various classifier for character recognition process of numerous Tamil scripts it is also proposed a noise image and segmentation process for individual characters image of letters between them. It also proposes to use advanced method for pre-processing and segmentation. It involves the use of proposed methodology called significant level of probability error to improve the images image quality for blurry images. It also describes about converting the image into a text format using ASCII code and other method is by using Matlab supports standard data and image format exchange.

Inference:

It is important to note in this paper that it proposed additional technique called significance level of probability error for blurry images to improve the image quality. Apart from the traditional methods for character recognition this paper proposes an additional technique which might cause computational overhead to the system. At the same time, it has also introduced ASCII code for converting the image into text format.

4. Tamil language character processing steps:

The process of converting scanned images of machine printed or handwritten text into editable text is called as Optical character recognition (OCR). In this process, OCR is developed for Tamil language.

Any character recognition process goes under the following steps[6][7]:

- Preprocessing
- Segmentation
- Feature Extraction
- Classification

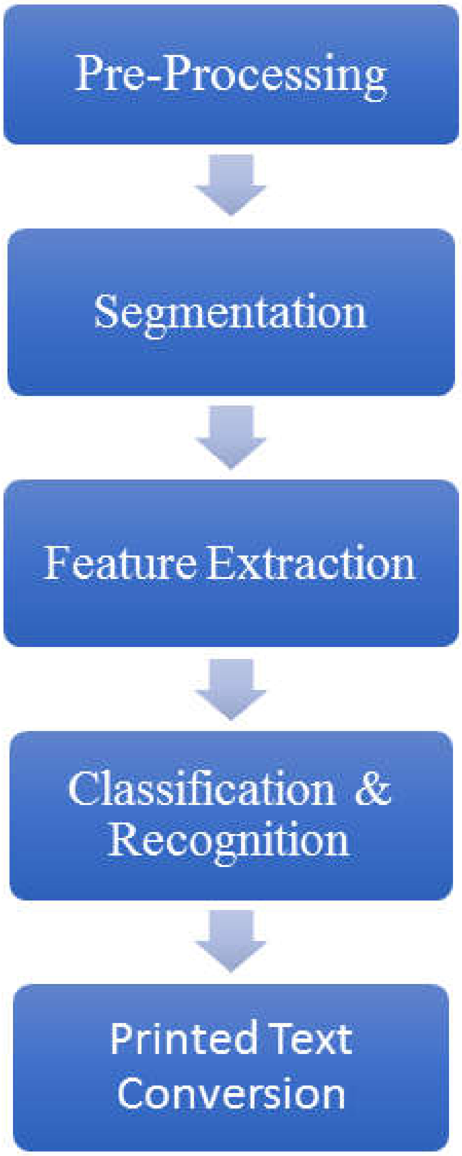


Figure1:Fundamental Tamil language character digital processing steps

The proposed system process helps people belonging to different arenas of work. This process involves capturing the image of handwritten Tamil Text and reproducing it in a computerized text format. Captured Image has to run through the proposed software which will help the image to get converted in digital text format. For Character Mapping, the American Standard Code for Information Interchange (ASCII) values can be recognized the Tamil characters and each of the fonts is matched with its corresponding template converted and saved as normalized text transcription languages. This proposed works offline which is done for the 1st time for Tamil Language. We finished our work on primary vowels like அ,இஉ,எ,ஒ& Tamil Consonant letters like □□,□□,□□,□□,□□ and its few compound forms.

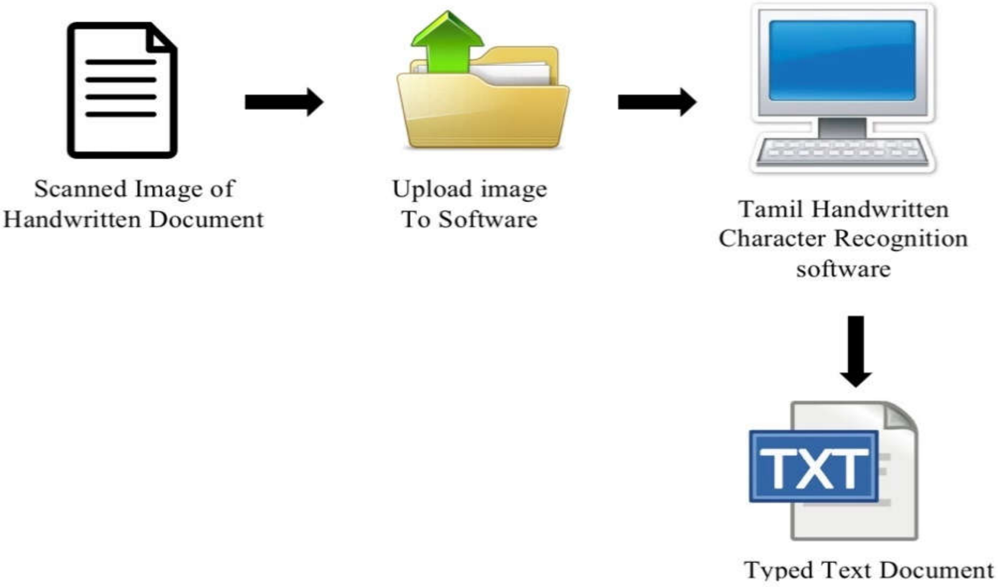


Figure 2: System Representation

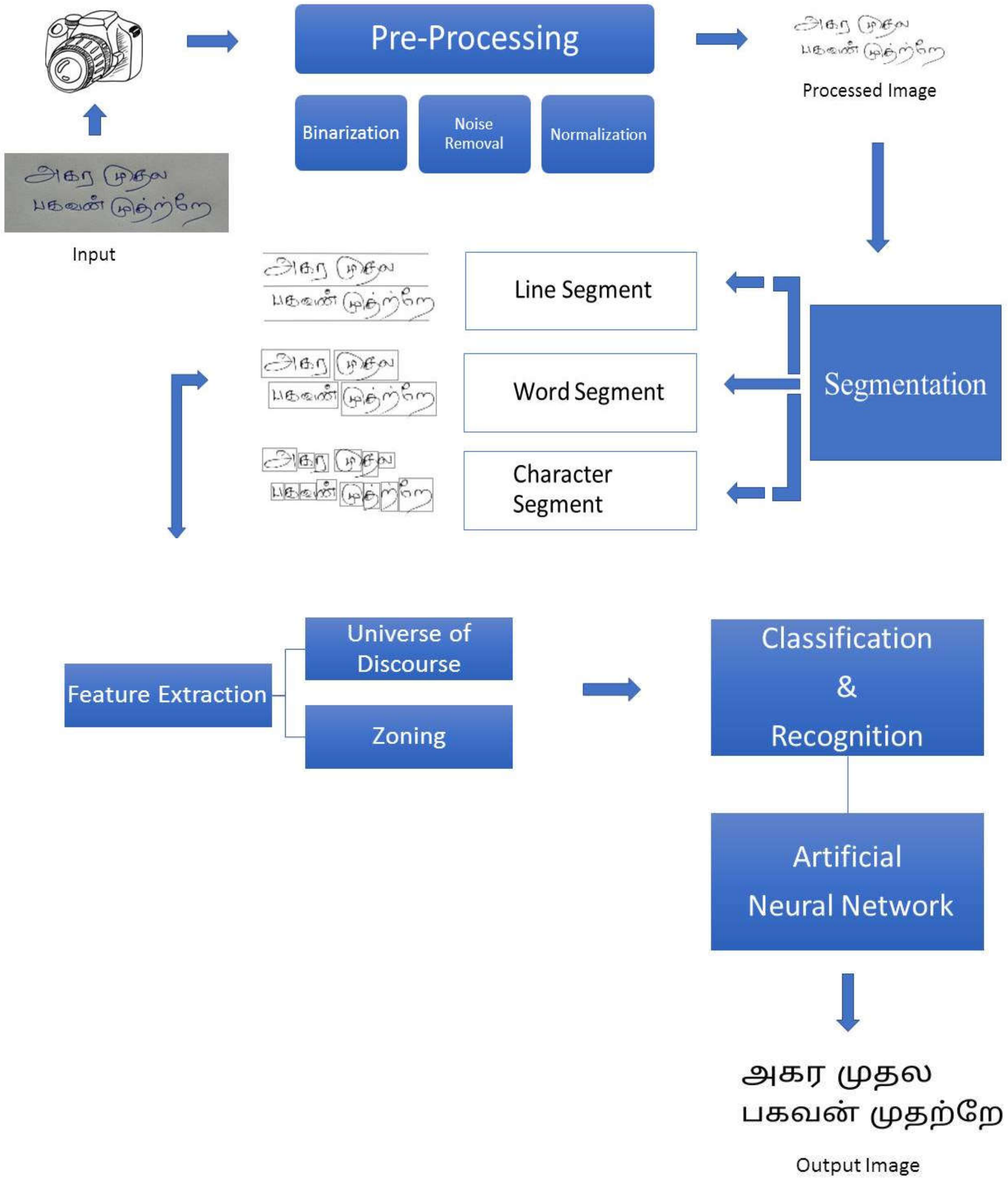


Figure 3: Tamil Language Character Digital conversion steps

The system is designed in such a way that it recognizes the handwritten text. In order to satisfy these criteria, the process involves inputting the Image of the Handwrittendocument into the software where it undergoes several procedures such as Pre-processing, Segmentation, Feature Extraction, Classification & Recognition and finally Post processing is done[6]. Generally, it involves conversion of the handwritten text of Tamil language which is recognized by using the Neural Network for Classification and Recognition step then the recognized characters are converted to typed text by performing ASCII Character Mapping. Finally, the recognized character is displayed in the screen as the document.

5. Training:

To train this classifier, an arbitrary number of examples for each letter are taken to be the training set for the letter. Each example of this letter's training set is given to a function. This function takes in an image and a letter. The function is run on all of the examples in the training set, and logs the vector obtained from the image into a cluster of vectors. This cluster of vectors describes all of the examples in the training set for some letter, and is written into a configuration text file. Each cluster of vectors is analyzed to find a mean vector for the particular letter. To retrieve the identity of a new image we take the distance from each mean vector to the input image vector (as described above). A distance is computed to each mean-vector of the various letters. The letter that yields the smallest distance is used to classify the image. If the smallest distance exceeds a threshold, then the image is declared to belong outside the training set.

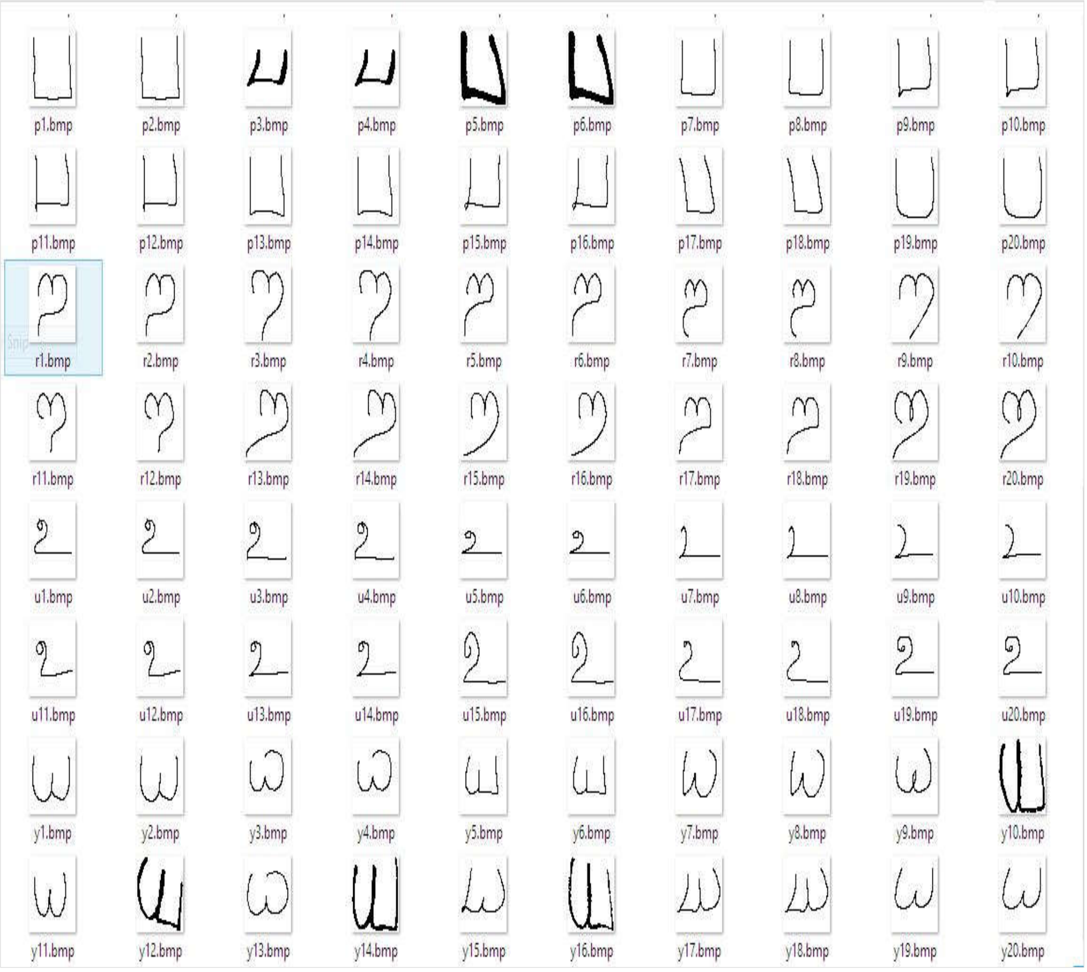

p1.bmp p2.bmp p3.bmp p4.bmp p5.bmp p6.bmp p7.bmp p8.bmp p9.bmp p10.bmp
p11.bmp p12.bmp p13.bmp p14.bmp p15.bmp p16.bmp p17.bmp p18.bmp p19.bmp p20.bmp
r1.bmp r2.bmp r3.bmp r4.bmp r5.bmp r6.bmp r7.bmp r8.bmp r9.bmp r10.bmp
r11.bmp r12.bmp r13.bmp r14.bmp r15.bmp r16.bmp r17.bmp r18.bmp r19.bmp r20.bmp
u1.bmp u2.bmp u3.bmp u4.bmp u5.bmp u6.bmp u7.bmp u8.bmp u9.bmp u10.bmp
u11.bmp u12.bmp u13.bmp u14.bmp u15.bmp u16.bmp u17.bmp u18.bmp u19.bmp u20.bmp
y1.bmp y2.bmp y3.bmp y4.bmp y5.bmp y6.bmp y7.bmp y8.bmp y9.bmp y10.bmp
y11.bmp y12.bmp y13.bmp y14.bmp y15.bmp y16.bmp y17.bmp y18.bmp y19.bmp y20.bmp

Figure4: Collected Data Sets

6. Handwritten Text To Typed Text :

The conversion of Handwritten text to typed text is done by using the American Standard Code of Information Interchange (ASCII) Character Mapping Technique. In this Technique, each recognized character is assumed to belong any one of the Tamil character obtained as the result form Neural network. The resultant character is recognized and identified to be any one of the Tamil Character set given. The ACSII character code for typed text is already made to be mapped with the respective Tamil character set. The identified character is converted from Handwritten form to Typed text form. This process is done serially for all recognized characters. So finally, we obtain the conversion Handwritten Text format to Typed Text format.

For Example, let us consider the character in the image is recognized and identified to be "Ra" (ற) Then ASCII value for character is "0BB1" which is mapped with recognized character to get the resultant output.

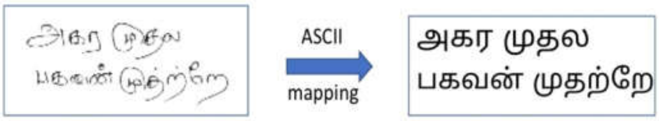


Figure 5: Example of ASCII character mapping for Conversion

7.Implementation Results:

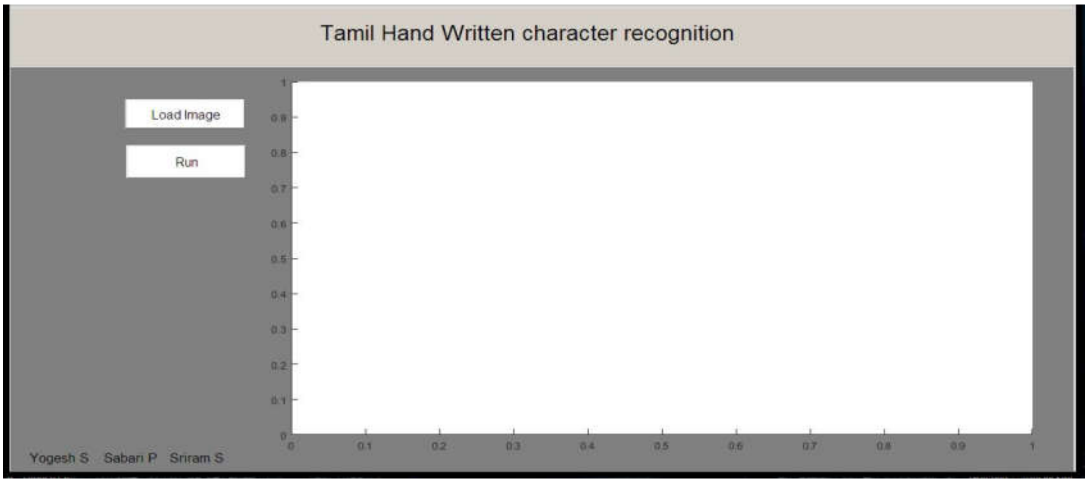


Figure 6: Home Screen of software

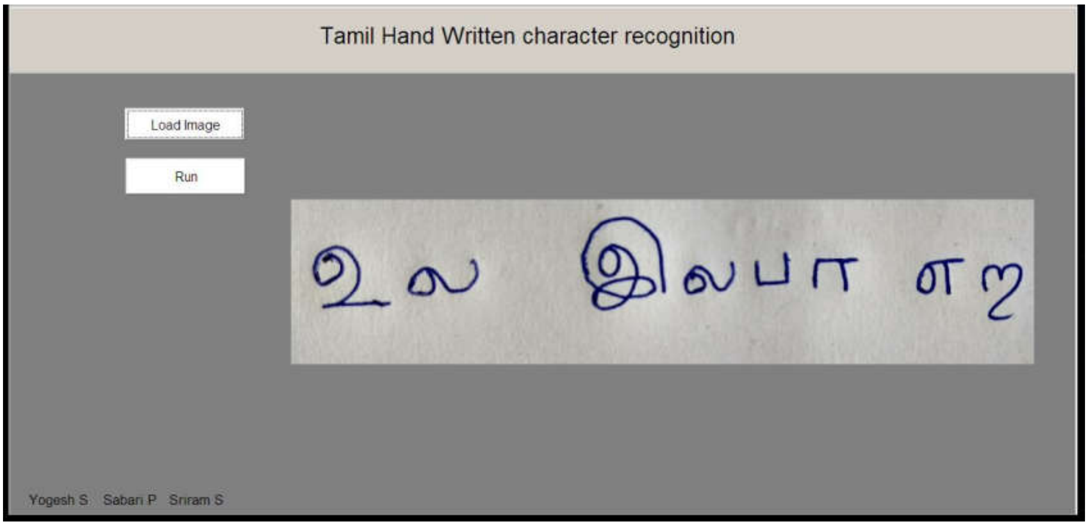
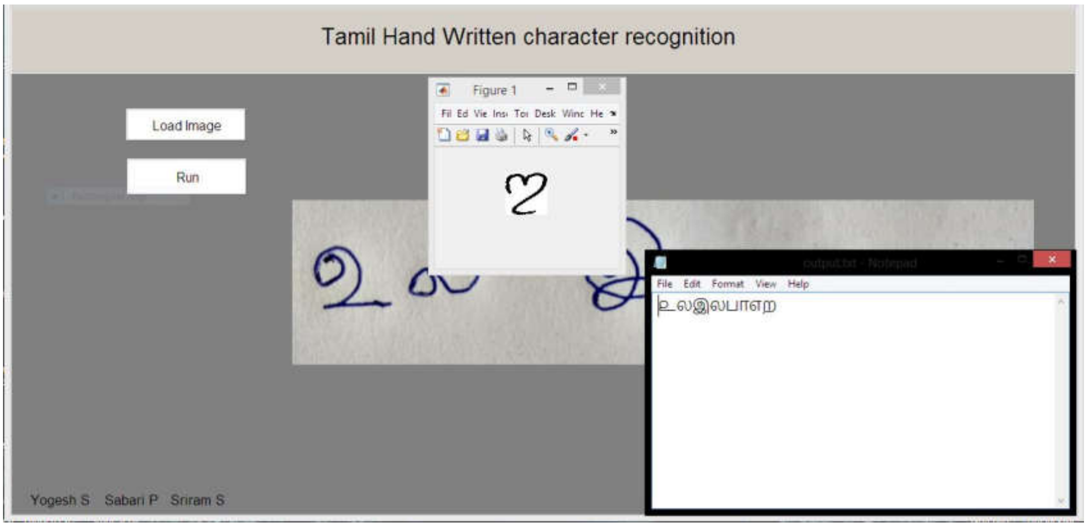

Tamil Hand Written character recognition
Load Image
Run
Yogesh S Sabari P Sriram S

Figure 7: Loading image


Tamil Hand Written character recognition
Load Image
Run
Figure 1
output.txt - Notepad
File Edit Format View Help
Yogesh S Sabari P Sriram S

```
உலஇலபாஎற
```

Figure 8: Resultant Output - Document Displayed

8.Conclusion:

In this project, a prototype system is developed which converts the handwritten character to typed text for seamless conversion of Handwritten document to Typed documents with ease. The project presents a complete Optical Character Recognition (OCR) technique followed by

Handwritten To Typed Text conversion. Various algorithms for optical character recognition have been studied and analysed. Based on the analysis the best algorithms are chosen and implemented in this project to make the system efficient. The advantage of this prototype is that it performs the handwritten character recognition of Tamil language in a single system. There is no implementation for Tamil language in the existing system. Hence this system focuses on developing libraries for few Tamil characters. Presently 30 characters are trained.

The feature work will be focused on training more Tamil characters and accuracy of the system will be enhanced. The system would be incorporated with mobile phones to keep it as simple as possible and economical along with operational feasibility.

References:

[1] https://sisu.ut.ee/imageprocessing/book/1

[2] http://www.onlineclassnotes.com/2011/10/describe-fundamental-steps-of-digital.html/

[3] K. Punitharaja and P. Elango," Tamil Handwritten Character Recognition: Progress and Challenges", 2014, International Science Press.

[4] D. Rajalakshmi and S. K. Jayanthi, " Extraction of Tamil character from a handwritten document using connected component labelling", 2017, International Journal of Computer Sciences and Engineering, Vol 5(9), E-ISSN:2347-2693.

[5] Surya Kala and Dr P Thangaraj ," Identification of Tamil Character Recognition by using MATLAB",2017, International Journal of Engineering and Technology(IJET) ISSN:2319-8613

[6] Daniel Keysers, Thomas Deselaers, Henry A.Rowley, Li-Lun Wang, and Victor Carbune, Multi-Language Online Handwriting Recognition, 2017, IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol.39,No.6

[7] Gauri Katiyar and Shabanab Mehfuz, " A Hybrid Recognition System for Offline Handwritten characters", 2016