Software for Tamil Handwritten Character Recognition (TCR)

P. Sabari[1] K. Shanmugam[2]
[1]B.E Student [2]Assistant Professor
[1,2]SRM Valliammai Engineering College, Chennai, India

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. The conversion of handwritten script to computerized text format is done. 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. 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 is designed to take input samples and classify them into groups as it is trained for classification. 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. This type is first of its attempt in Tamil language technology integration.

Key words: Network, Pre-processing, Segmentation, Feature Extraction, Classification, Normalization

I. 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 facilitate researchers and users to explore the language further. The major components of integration is image processing.

II. IMAGE PROCESSING

A physical procedure used to adapt an image signal into a physical image is called image processing. 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. Analog image creation is an older method where an image is captured into a film and is later processed using a chemical reaction which is triggered by controlled exposure of light. This type of image of processed in dark room with help of special chemical to make the actual image. the fact is that digital image us have taken over the analog images due to the advancement like effortless picture taking easy processing economical storage etc. the arena of digital image processing has given paved way for new range of applications and tools which dint exist previously. Image processing has created a revolution just not in photography filed but also in electronics gadgets like face recognition software, medical images like advanced scans and remote sensing like in weather as well. Each day computer programs are created to enhance and update image processing.

A. Digital image processing

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

III. OBJECTIVE

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.

IV. RELATED WORKS

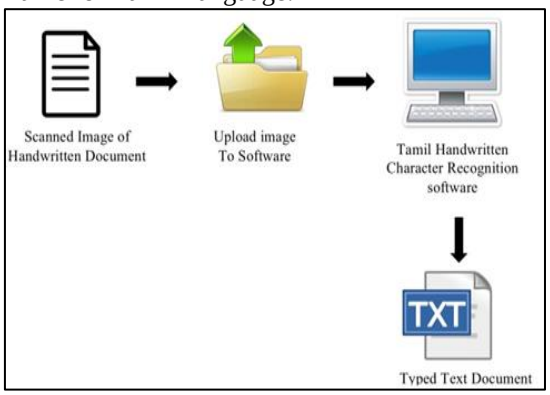
The research paper "Tamil Handwritten Character Recognition: Progress and Challenges " byK. Punitharaja and P. Elango (2014) states about Tamil character recognition system development, Attributes which evaluates techniques which undergoes in in Tamil character recognition. This paper debates about identifying 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.

The research paper "Extraction of Tamil character from a handwritten document using connected component labelling" by D. Rajalakshmi and S. K. Jayanthi (2017) discusses about 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.

The research paper "An Optical Character Recognition System for Tamil news print" by K H Aparna, Sumanth Jagannath, P Krishn Kumar, V S Chakravarthy (2009) is written in view with an early version of complete optical character recognition OCR system for Tamil news print. All standard elements of OCR are implemented in this paper. It uses the artificial neural networks ability to learn the obituary input-output mappings from the sample data for solving the key problems of segmentation and character recognition further the problems that are present in segmenting text are discussed.

V. PROPOSED SYSTEM

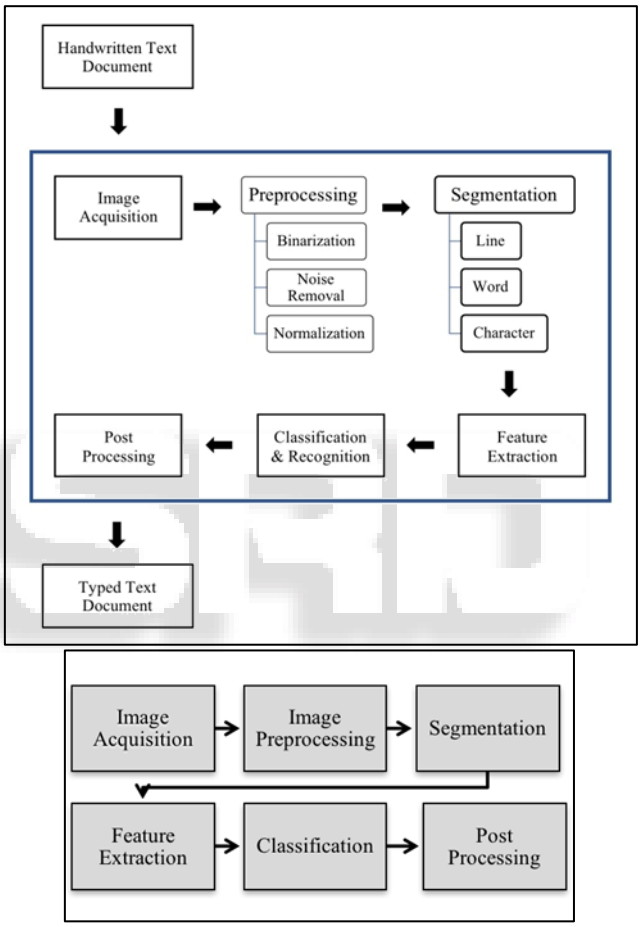
The proposed process helps people belonging to different arenas of work. This process involves capturing the image of handwritten Tamil Text and reproducing it in a computerized text format. Captured Image has to run through the proposed software which will help the image to get converted in digital text format. The handwritten character 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.



System Representation

VI. ARCHITECTURE DIAGRAM

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 Handwritten document into the software where it undergoes several procedures such as Pre-processing, Segmentation, Feature Extraction, Classification & Recognition and finally Post processing is done. 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.



Architecture Diagram of the System

VII. IMPLEMENTATION

A. *Project Modules:*

- Image Acquisition
- Character recognition
- Image Pre-processing
- Image Segmentation
- Feature Extraction
- Classification
- Post Processing
- Handwritten to computerized text format.

B. *Fundamental Steps involved in the process*

1) Image Acquisition:

Digital Image is fed into the software as the input. The Images could be a scanned copy through scanner or user can

manually browse and locate the prescanned images from the drive and feed it in the software. Images can be in any format such as .JPEG, .PNG, .BMP, .TIFF etc. are accepted by the software.

The challenges faced by the software was that due to low quality image of hand written script the improper response was the output by the software. So, it is important to have a high-quality hand-written image. Acquisition must be proper to produce desired output.

2) Character Recognition:

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.

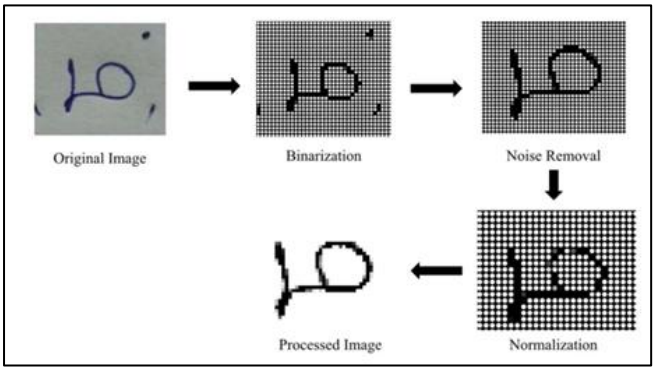
Any character recognition process goes under the following steps:

- Preprocessing
- Segmentation
- Feature Extraction
- Classification

3) Preprocessing:

Multiple procedures are involved in enhancing the image to make it suitable for segmentation. Unwanted noise in the image processing is removed by implementing suitable threshold morphological transformations has dilation and Binarization transformation, which generates the required contours and draws the bounding boxes around the required text.Proper filters like mean filter, min-max filter, Gaussian filter etc. can also be practical solutions to remove noise from document or image. Binarization is done using the inbuilt Matlab.

Binary Morphological operations like opening, closing thinning, hole filling etc. are applied to augment visibility and structural information of character. Input document may be resized if it is too large in size to reduce dimensions to improve speed of processing. However, reducing dimension below certain level may remove some useful features too.



C. Example of Pre-processing illustration-sample

1) Normalization:

Normalization is used in digital signal processing.The intensity value of the pixel to the range of 0,1 changes through process which is called normalization in image processing. The process which involves the conversion of various dimensional images into a fixed dimension is also known as normalization. Further, complications during feature extraction are removed if normalization is done in the earlier stages.

The intensity value of the pixel to the range of 0,1 changes through process which is called normalization in image processing.

2) Sampling:

Discretization of analog signal is called sampling. Pixel is the smallest element resulted out of discretization.The process of selecting the subset of individuals from the large sample of population and examining those samples, can generalize the results to the whole population.

3) Noise filtering:

While an image, scanned copy or digital copy directly from the drive,fed to the software, may or may not contain noise. If its presented with noise then noise filter is applied to remove the same. If the image contains no noise, the application may not be used at all here.

4) Thinning:

The pre-process of specific pixel width image to identify the handwritten character is called Thinning. It is applied repeatedly leaving only pixel-wide linear representations of the image characters.

5) Segmentation

Segmentation is the process for isolating words into specific characters. The first process on preprocessed image is segmentation where we use the following algorithm to extract all characters from the image. Segmentation was performed in two phases:

6) Line Segmentation

When the image seems ready for processing, every line of the image is segregated. The image skimmed horizontally through a computer program in line segmentation. The selected region represents the line that holds a single or more characters. Then the whole image and each recognized line is saved in a short-term array for more image processing.

7) Character Segmentation

Character segmentation is one another technique through which the feature extraction is used to isolate or detect characters from digital images. Characters are isolated and detected through scanning every array vertically within every line after going through line segmentation. The character borders are the beginning and the black pixels, which are vertically detected. At the same time, as the edges of every character box are needed for recognition purposes, another scan is performed horizontally to detect the lower and the upper end of the character and sequestrate the region which consists only of character pixels.

8) Feature Extraction

Feature extraction is the heart of any pattern recognition application. The characters features that are considered vital for categorising them at the stage of recognition are obtained in this stage. Feature Extraction is a method that is able to detect straight lines, curves or any particular shapes. Techniques involved are Zoning method. This technique might be applied to extract the features of individual characters which are used to train the system.

9) Classification

Classifiers like artificial neural network compare the input feature with stored pattern and find the best matching. Factors affecting the development of Tamil character Recognition (TCR) system are random factors and linguistic factors. Random factors affect the document scanning process. Example: Ink and dirt spattering, paper quality,

quality of writing tools. Linguistic factors are the intrinsic part of the Tamil language. The cardinality of the Tamil alphabet set is one of the linguistic factors that affect the design of TCR system.

The purpose of the classifier is to take a 50x50 pixel image and classify it as a letter in the Tamil alphabet. This particular classifier needs to be trained on each letter that will be recognized. This process requires 50x50 pixel images. It uses back propagation techniques and ideas taken from artificial neural network. An array of some attributes is interpreted as a vector, in the mathematical sense. This vector has a dimensionality that varies with the number of attributes garnered from the image.

10) Post Processing

The conversion of Handwritten text to typed text is done. 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 identified character is converted from Handwritten form to Typed text form. This process is done serially for all recognized characters.

VIII. 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. So finally, we obtain the conversion Handwritten Text format to Typed Text format.

REFERENCES

[1] K.Punitharaja and P. Elango,"Tamil Handwritten Character Recognition: Progress and Challenges", 2014, International Science Press.
[2] 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.
[3] K H Aparna, Sumanth Jagannath, P Krishna Kumar, V S Chakravarthy,"An Optical Character Recognition System for Tamil newsprint", 2009.
[4] Aparna K G and A G Ramakrishnan ,"A Complete Tamil Optical Character Recognition System", 2001.
[5] Surya Kala and Dr P Thangaraj ,"Identification of Tamil Character Recognition by using MATLAB",2017, International Journal of Engineering and Technology(IJET) ISSN:2319-86

