INTELLIGENCE SYSTEM FOR TAMIL VATTEZHUTTUOPTICAL CHARACTER RCOGNITION

Mr R.Vinoth
Assistant Professor, Department of Information Technology
Agni college of Technology, Chennai, India
vinoth.it@act.edu.in

Rajesh R.
UG Student, Department of Information Technology
Agni college of Technology, Chennai, India
rajesh110595@gmail.com

Yoganandhan P.
UG Student, Department of Information Technology
Agni college of Technology, Chennai, India
yoganandhan94@gmail.com

***Abstract*--A system that involves character recognition and information retrieval of Palm Leaf Manuscript. The conversion of ancient Tamil to the present Tamil digital text format. Various algorithms were used to find the OCR for different languages, Ancient letter conversion still possess a big challenge. Because Image recognition technology has reached near-perfection when it comes to scanning Tamil text. The proposed system overcomes such a situation by converting all the palm manuscripts into Tamil digital text format. Though the Tamil scripts are difficult to understand. We are using this approach to solve the existing problems and convert it to Tamil digital text.**

Keyword - Vatteluttu Tamil (VT); Data set; Character recognition; Neural Network.

I. INTRODUCTION

Tamil language is one of the longest surviving classical languages in the world. Tamilnadu is a place, where the Palm Leaf Manuscript has been preserved. There are some difficulties to preserve the Palm Leaf Manuscript. So, we need to preserve the Palm Leaf Manuscript by converting to the form of digital text format. Computers and Smart devices are used by mostof them now a day. So, this system helps to convert and preserve in a fine manner.

A system that scans the Palm Leaf Manuscript and stores in the form of image format and then converts them into a digital text format. First thing is to convert the Palm Leaf Manuscript into image format by scanning or through snapshot and to store it in a database. After that stored images should be converted into digital text format.

II. OBJECTIVE

The main objective of the system is to make the Script of Tamil language more accessible. This system can be used in a variety of ways depending on the requirement of the user. The students may not understand the ancient Tamil language what they see in Palm leaf manuscript. But this System does. And this will help them to a greater extent and will allow them to be independent.Given a Palm leaf manuscript of Tamil to any person with the basic knowledge of Tamil language and that may or may not understand the content, by converting it to the present Tamil language. This system can also be useful for archiving purposes. It is helpful in digitizing the Palm leaf manuscript by converting and storing it in a digital format.

III. ANCIENT LANGUAGE

1. Brahmi
2. Vatteluttu
3. Pallava
4. Grantha

A. Brahmi

It is one of the important writing systems in the world. It represents the earlier texts found in India. The bestknown Brahmi inscriptions are the rock-cut edicts of Ashoka in North-central India, dated to 250–232 BCE. This elegant script appeared in India most certainly by the 5th century BCE, but the fact is that it had many local variants even in the early texts which suggest that its origin lies further back in time.

B. Vatteluttu

The Vatteluttu (or Vattezhuttu) script was the one used mostly in the southern part India. It originated from the southern form of Brahmi script during 6th century and then it was permitted to write Tamil and Malayalam languages. It inherited from Brahmi to state that the Tamil language was eliminated from the Vatteluttu script. A feature also found in the Tamil script and most likely was an influence from Vatteluttu.

C. Grantha

The Grantha alphabet is a type of the earlier language related to the Brahmi alphabet and started emerging during the 5th century AD. Most of the alphabets of southern India originated from Grantha and it also used highly in Sinhala and Thai alphabets. The Grantha alphabet has been traditionally used by Tamilians to write Sanskrit and is still used in traditional Vedic schools.

D. Pallava

The Pallava script was originated in southern India during the Pallava dynasty. The Pallava script was based on the Brahmscript. At first the script was used to write Sanskrit, after that used in various languages. Later it became popular in religious and inscriptions on stone monuments and it was used for nearly 500 years. other scripts developed from Pallava and many other languages including Telugu, Kannada, Tamil, and Malayalam. The script is also known as Southern Gupta Brahmi, Tamil Grantham.

IV. LITERATURE REVIEW

Historically, handwritten character recognition applications use three major approaches: statistical approach, structural or syntactic approach and neural network-based approach.

A. Statistical Approach

Statistical pattern recognition uses statistical or probability functions for building a recognition algorithm. The input features are extracted from a set of characteristic pattern measurements. A limitation of this approach is the difficulty that exists in expressing pattern classification in terms of structural information.

B. Structural or Syntactic Approach

Structural approach uses structured information about how to generate knowledge. This approach checks the similarities between patterns and generates syntax for structural rules. The information about pattern syntax for structural rules is used to identify and analyse unknown patterns. This approach is suitable for converting a handwritten character recognition system because it uses structural approach to convert unlimited handwritten character pattern syntax.

C. Neural Network Based Approach

Neural Network approach states that how the neural system stores and retrieves the information. This system is called Neural Networks. The function of the neural network is to solve all the problems in an automatically. This also includes a pattern recognition problem. This approach gives clear idea about how the patterns are used to predict the properties of neural networks.

V. CONVERSION METHOD

In this systemfirst scans the palm leaf manuscript and then uploading into the system then using the image processing and converting into the digital format and storing into the system that shown in figure-1.

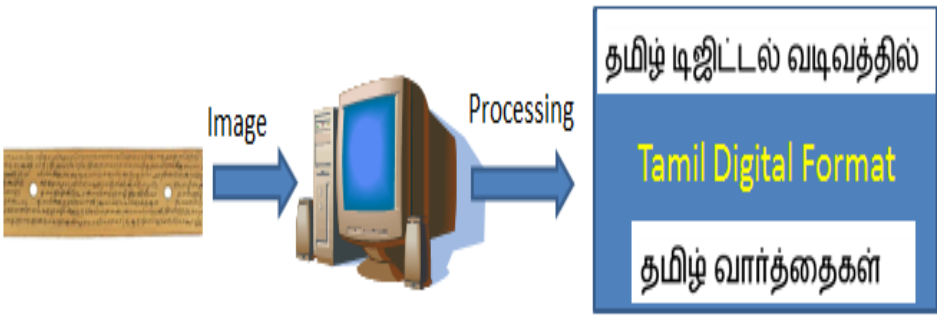
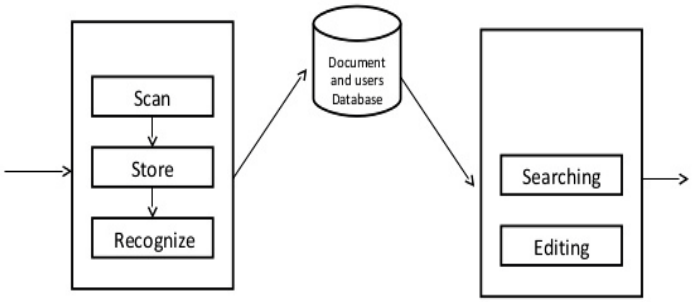


Figure-1 Architecture of Conversion of Tamil palm leaf to digital format.

A. System architecture

The figure-1 shows that the conversion of Tamil palm leaf manuscript into the Tamil digital-format



OCR SYSTEM ARCHITECTURE

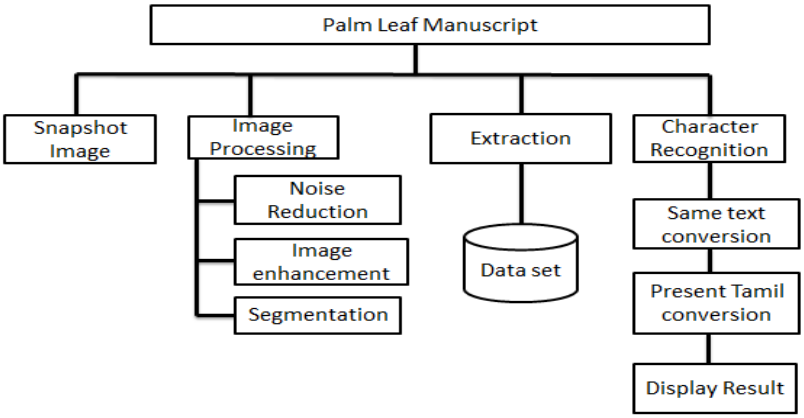
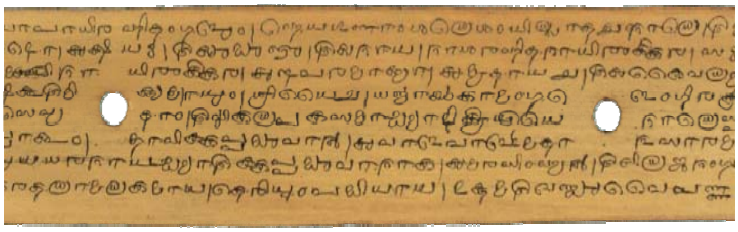


Figure-2 The Process of conversion

The figure-2 shows the conversion of Tamil palm leaf manuscript to digital format. There are three frame work for conversion (i) Snap shot (ii) Image processing (iii) Extraction (iv) Character recognition

B. Snap shot (Image)

Infirst stage the palm leaf manuscript collected as an image.The snapshot images are captured by high quality or high resolution HD / DSLR (High Definition / Digital Single Lens Reflex) camera and stored in Jpeg format.



C. Image Processing

In the image preprocessing module, the system prepares a character image for the feature extraction. This stage consists of three subprocesses: (i) Noise Reduction (ii) Image Enhancement (iii) Segmentation

a) Noise Reduction:

Noise is the result of errors in the image acquisition process that result in pixel values that do not reflect the true intensities of the real scene.

b) Image Enhancement:

Image enhancement techniques bring out the Information of an image highlightscertain features of an image enhancement techniques. Which include contrast adjustment, noise filtering, morphing, and deblurring. An image enhancement modifies the original image.

c) Segmentation:

Segmentation is nothing but partitioning the image into distinct regions. Which is used for image analysis and interpretation of the regions should be strongly related to the features interested. Segmentation is the process of converting grey scale or colour image into one or more other images into high-level in terms of features, objects, and scenes. The success of image analysis depends on reliability of segmentation.

D. Extraction

The extraction is the process where it extracts the data from the database

a) Data Set:

In this stage the dataare stored and the data are the Tamil letters, from the data set the characters are extracted.

E. Character Recognition

In this stage it will recognition the character from the palm leaf manuscript letters and compare with the stored letters.

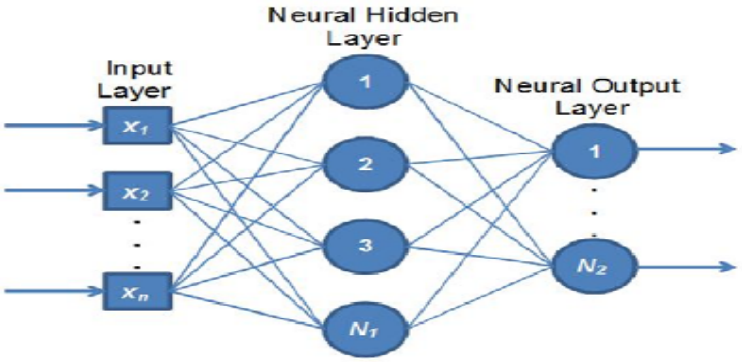
a) Conversion of same Tamil letter

In this stage the palm leaf manuscript letters will be converted into digital format of the same palm leaf manuscript.

F. Conversion of Present Tamil letter

In this stage the palm leaf manuscript letters will be converted into the present Tamil character.

VI. PROPOSED ALGORITHM



A. Input Layer

Input layer takes input from the external world and encodes it to the convenient formin the input layer. Which is called conduit through which the external environment provides the neural network. Each and every input neuron should represent some independent variable that can able to generatethe output of the neural network. The input neurons OCR is calculated using the number of pixels in the image.Characters of the image must be represented by a 5×7 format with 35 pixels. So, that it will generate 35 input neurons.

B. Hidden layer

Hidden layer cannot see or act upon the outside world directly. These interneurons communicates only with other neurons. Deciding the number of neurons in the hidden layers is very important part of deciding your overall neural network architecture. Both hidden layers and neurons must be carefully considered. There are many rules of thumb impression detection methods for determining the exact number of neurons to be uses in the hidden layers such as follows:

• The number of hidden neurons should be ranged between the size of the input layer and the size of the output layer.

• The number of hidden neurons should be ranged as 2/3 for considering it as the size of the input layer and the size of the output layer.

• The number of hidden neurons should be less than the size of the input layer.

C. Output Layer

The output layer of the neural network provides a pattern to the external environment. The number of output neurons should be directly related to the work to be performed. The number of output neurons used by the OCR will differ as per the characters contained in a program. The default file that is provided with the OCR is used to recognize 256 characters. Using this, the neural network will generate output as 256 neurons. Providing input to the neuron should correspond to the output neurons in such a way the input information should be given. After receiving the required input information the neural network will direct us to the exact output neurons.

VII. CONCLUSION

The system aims to convert the palm leaf manuscript into the digital format. This has been done for storing the palm leaf manuscript information because the palm leaf manuscript were difficulties for preserving and storing and using this we can store it in system if the palm leaf manuscript were destroyed. The upcoming update of this system will be for other Tamil languages.

REFFERENCE

[1] Giridharan.R, Vellingiriraj.E.K, Dr. Balasubramanie.P, "Identification of Tamil ancient characters and information retrieval from temple Epigraphy using image zoning" ICRTIT 2016 International Conference on recent trend in Information Technology.

[2] Chamila Liyanage; Thilini Nadungodage; Ruvan Weerasinghe "Developing a commercial grade Tamil OCR for recognizing font and size independent text" 2015 15th International Conference on advanced in ICT for emerging region.

[3] Honey Mehta, Sanjay Singla, Aarti Mahajan "OpticalCharacter Recognition (OCR) System for Roman Script and English language using artificial Neural Network" 2016 International Conference on Research advanced in Integrated Navigation System (RAINS).

[4] Pelin Gorgel, Oguzhan Oztas "Recognition of HandwrittenCharacter using Neural Network" International Journal of innovative research in Computer and Communication Engineering. June 2016 Vol-4 issue-6.

[5] Mohamed Sageer T. K.* Dr. A.T. Francis "Palm leaves manuscripts in kerala and their preservation: factors necessitating digital archiving", Research Scholar, Department of Library and Information Science, Karpagam University, Coimbatore,

[6] Sudarshan Sawant, Prof. Seema Baji "Handwritten character and word recognition using their geometrical features through neural networks" Department of electronics and Tele-communications, Late G N Sapkal College of Engineering, Nasik, Maharashtra.

[7] J.Pradeep, E.Srinivasanand S.Himavathi "diagonal based feature extraction for handwritten alphabets recognition system using neural network"Department of ECE, Pondicherry College Engineering, Pondicherry, India. jayabala.pradeep@pec.edu