ISSN:0975-9646

A Survey of Methods and Strategies in Tamil Palm Script Recognition

B. Kiruba[1], Mr.K.Manikandan[2]
PG scholar, Department of Computer Science and Engineering [1]
Sri Krishna College of Engineering and Technology, Coimbatore.

Assistant Professor, Department of Computer Science and Engineering[2]
Sri Krishna College of Engineering and Technology, Coimbatore.

Abstract - Palm leaf manuscript is the conventional medium of writing in Asian Countries. In order to recognize palm manuscript character, few research works were carried out on Indian languages like Devanagari, Kannada, Telugu scripts and now, various researches are in the progress for the development of efficient system for recognizing the Tamil language. The ancient Tamil literature includes medical works, astronomy & astrology, Siddha and so on. This work describes the survey of various techniques being used for Tamil palm script recognition.

Keywords: Optical Character Recognition, Tamil Palm Script, Connected Component, Projection Profile, Genetic Algorithm, FFNBP.

I. INTRODUCTION

Optical Character Recognition is an approach that implements a conversion of different types of documents such as scanned papers. This technique can be categorized as either offline or online. In off-line handwriting recognition, the writing is usually acquired optically by a scanner and complete writing is available as an image. The online handwriting recognition is also referred to as real-time recognition because here the characters are recognized as they are written. Online HCR involves the use of pen based input devices to capture the sequence of co-ordinate points as the character is written. Recognition of handwritten character gets complicated due to numerous variations involved in the shape of characters, different writing style, overlapping and the interconnection of the neighboring characters, It also depends on the individual since we do not write the same character in that same way, Since developing an OCR system with high recognition accuracy for Tamil script is a difficult task. The main objective of this system is to recognize Tamil characters present in palm manuscripts. This is done by classifying the characters into appropriate types based on features extracted from each character. Following steps aid in acquiring high accuracy.

1. Image Preprocessing
2. Segmentation
3. Feature Extraction and selection
4. Classification

The organization of the paper is as follows. Characteristics of Tamil Script are described in section II. An Architecture diagram is given in section III. Pre-processing techniques are surveyed in section IV. Section V describes various segmentation methods presented for offline character recognition. Feature extraction methods are discussed in section VI. Various classification approaches are explained in Section VII followed by conclusion and references in Section VIII and IX respectively.

II. CHARACTERSTICS OF TAMIL SCRIPT

Tamil is a traditional language which is broadly spoken in most part of the south India. The unique Tamil script consists of 12 vowels, 18 consonants and one special character. Each pure consonant can combine with each vowel to produce a sum of 216 consonant-vowel (CV) combinations. These add up to a total of 247 Tamil characters. Alphabet system of Tamil language is assumed to be a derivation from the prehistoric Brahmi script which serves as a foundation for most of the Indian languages. The vowels and consonants of Tamil alphabet set are given in the table 1.1:

Table 1.1 Modern Tamil character set

Vowels	அ, ஆ, இ, ஈ, உ , ஊ ,எ , ஏ , ஐ , ஒ , ஓ , ஔ
Constants	க், ங் , ச் , ஞ் , ட் , ண் ,த் , ந் , ப் ,ம் , ய் ,ர் ,ல் , வ் ,ழ் , ள் L, ற் , ன்
Grantha	க்ஷ, வ்ஷ, ஸ்ரீ
Aytam	ஃ

In Tamil a well developed handwritten Tamil character Recognition system is still not available. The main reasons for this are:

- Tamil Language has a very large character set
- Due to complex letter structure, writing styles of people vary significantly
- There is no Tamil character database that exists for testing purposes in the public domain.

III. ARCHITECTURE DIAGRAM

The Tamil Palm manuscript recognition process is shown below in the figure 3.1. A palm manuscript is chosen for scanning. The Scanned image is preprocessed and lines, words, characters are segmented. Features are extracted from the segmented characters. Then recognition can be achieved by classification of characters.

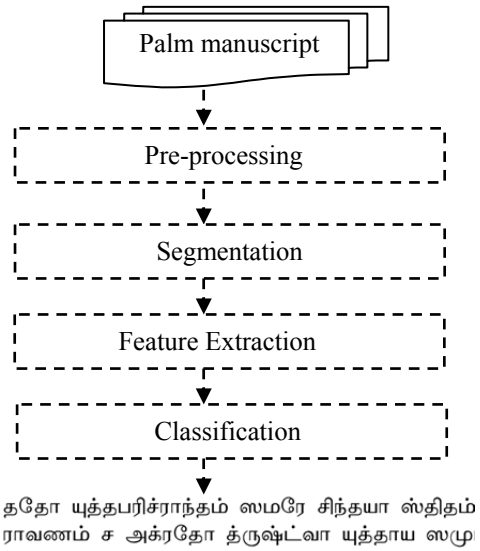

Fig 3.1 Tamil Palm Manuscript Recongition Process

IV. PREPROCESSING

Each time data is collected for recognition, it is collected as an optically scanned image of the paper document. Text is converting into digital form by using a flat bed scanner having resolution between 100 and 600 dpi and stored. These pixels may have values: 0 or 1 for binary images, 0–255 for gray-scale images and three channels of 0–255 color i.e. RGB values for color images. This collected raw data should be further analyzed to get useful information. Pre-processing essentially improves the image for suitable segmentation. Such processing includes the following:

A. RGB to Gray Conversion:

The scanned picture is stored as a JPEG image but images of other formats like BMP, TIFF etc are also used for recognition. All these images are in RGB format are converted into grayscale, then the RGB values for each pixel and make as output a distinct value reflecting the intensity of that pixel. One such approach is to take the average of the contribution from each channel: (R+B+C)/3.The value of a pixel lies under 0 to 1 or under 0 to 255 depending upon its class.

B. Thresholding/Binarization:

Binarization is a technique of converting a gray scale image into a binary image by using global thresholding technique. This image arrangement also stores an image as a matrix but can only color a pixel black or white. It assigns zero for black and one for white. Then it is inverted to obtain image such that object pixels are represented by 1 and background pixels by 0.

C. Noise Reduction:

The optical scanning device or the writing instrument introduced the noise which causes disconnected line segments, bumps and gaps in lines, filled loops etc. The distortion together with local variations, rounding of corners, dilation and erosion, is also a crisis. Median filter is a process that replaces the value of a pixel by the median of gray levels in the neighborhood of that pixel.

D. Skew Detection and correction

Skew Detection refers to the incline in the bitmapped image of the scanned picture. It is usually caused if the paper or palm script is not fed straight into the scanner. Many researchers proposed an algorithm to estimate the skew angle which is a position angle from the horizontal or vertical direction. To remove the skew present in image, the text is rotated into opposite direction. It must be a zero degree. The skew in the document images can be classified into three different types such as global skew, multiple skew and non uniform text line skew.

E. Thinning

Thinning is one of the morphological operation that is used to eliminate the chosen foreground pixels from the twofold images and skeletal the images to single-pixel width level so that their contours are brought out more intensely In this way, the attributes to be examined later and it will not be affected by the uneven thickness of edges or lines in the symbol. Different standard functions are now available in MATLAB for above operations.

V. SEGMENTATION

Segmentation is a significant step in recognition system as it extracts important regions for additional analysis. It generally contains the subsequent process [1]:

- First find the text lines in the page.
- Identify the words in individual line.
- Lastly, identify individual character in each word.

A. Connected Component Technique

The connected component technique initially labels the pixels in the image. The pixels that are connected are labeled with the similar blob [11]. This connectivity can be 4 or 8. After labeling, the labeled components are extracted from the image. The Connected Components process solves the overlapping character segmentation problem, but it separates the simple characters into their basic glyphs which may raise the recognition complexity [12].The characters are segmented into two glyphs each. These glyphs are to be reassembled to conserve the character contour if the recognition phase uses the shapes of the basic characters. Connected components grouped the scanned and also labeled its pixels into components based on pixel connectivity i.e. all pixels in a connected component divide related pixel intensity values and are in some way associated with each other. Afterwards all groups have been identified, each pixel is labeled with a gray level or a color according to the component it was selected. Connected component labeling works by scanning an image, pixel-by-pixel (from top to bottom and left to right) in order to classify connected pixel regions, that is regions of pixels which have the identical intensity values. The connected components labeling operator scans the image by moving along a row until it extends to a point *p* (where *p* denotes the pixel to be labeled at any stage in the

scanning process) for which *V*= {1}. When this is right, it inspects the four neighbors of p which have already been encountered in the scan. Based on this information, the labeling of p occurs as follows:

- If all four neighbors are 0, allocate a new label to *p,* else
- If only one neighbor has *V = {1}*, assign its label to *p*, else
- If excess of one neighbor has *V = {1}*, assign one of the labels to *p* and make a note of the equivalences.

After carrying out the scanning process, the pairs with the same labels are sorted into related classes and a special label is selected to each class. Lastly second scan is made through the image, during which every label is replaced by the label assigned to its related classes.

B. Projection Profile methods

A normal selection for line segmentation of gray scale images is the projection profile method [2]. Interspaces between the texts lines can be found by finding the greatest projections values, the projection value is intended by summing the pixel values by the side of the horizontal directions of the manuscript image. There are two main advantages for the projection profile approach in the perspective of historical document. To fragment the text lines, from the manuscript, the horizontal projection profile is calculated. The horizontal projection profile is the histogram of the number of depth values of the pixels along each row of the image. The space between text lines is used to section the text lines. The projection profile will have histogram of zero height between the text lines. Line segmentation is completed at these points [3].In order to fragment the word from the text line the vertical projection profile of an input text line is considered. Vertical projection profile is the summation of ON pixels along each column of the image which is used to take apart the word from the text line.

C. Particle Swarm Optimization

The distance between the lines is used to separate the lines. Generally the space between two lines are larger than the space between words, thus lines can be divided by comparing this distance next to a suitable threshold. To find out an optimal threshold, Particle Swarm Optimization method is used. It is known from literature, Particle Swarm Optimization (PSO) algorithm is used to resolve several difficult problems in the field of pattern recognition [13]. Hence, PSO is used to compute an optimal value.

Let *A* and *B* denote the particle's position and its corresponding velocity in search space respectively. Velocity and position of each particle in next iterations can be calculated using following equation (1.1) and (2.1)

$$B_{ij}^{k+1} = Wb_{ij}^{k} + c_1\, r_1\left(pbest_{ij}^{k} - A_{ij}^{k}\right) + c_2 r_2\left(gbest_{ij}^{k} - A_{ij}^{k}\right) \quad (1.1)$$

$$A_{ij}^{k+1} = A_g^k + B_g^k \quad (1.2)$$

where *k* is the current iteration number is inertia weight, *Bij* is then updated velocity on the *j*th dimension of the *i*th particle, *c1* and c2 are acceleration constants, *c1* and *c2* are positive constant parameters, usually *c1 = c2* = 2. *r1* and *r2* are the real numbers drawn from two uniform random sequences of *U (0, 1).*

PSO algorithm randomly generates the initial population of the PSO. In PSO, every particle is initialized with locations and velocities using equations (1.1) and (1.2) .These locations consist of the initial solutions for the optimal threshold. The process of the PSO algorithm is described as follows:

Step1: Initialize N particles with random positions.
Step2: Evaluate each particle by using equation (1.3)

$$f(t) = \omega_{0(t)} \times \omega_{1(t)} \times \left(\mu_{0(t)} - \mu_{1(t)}\right)^2 \quad (1.3)$$

Step3: Update individual and global best positions and also update velocity
Step4: Update velocity: update the *i*th particle velocity using the Eq. (1.2) restricted by maximum and minimum threshold vmax and vmin.
Step5: Update Position: update the *i*th particle position using Eq. (1.1) and (2.1).
Step6: Repeat step 2 to 5 until a given maximum number of iterations is achieved or the optimal solution so far has not been improved for a given number of iteration.

The greatest threshold value is obtained using PSO in order to divide text line from the image. To calculate the text line vertical projection profile, the word will be segmented. In the profile, the zero valley peaks may exhibit the character or word space. To characterize whether it is character or word spacing, find the maximum character space cluster and use it for separating the words.

D. Character Segmentation

Character segmentation from word is complicated because vowel modifiers placed on top of base characters and consonant modifiers close to left or right or bottom of the base character. Projection profile method does not give good results, when the characters are touched or overlapped. Therefore the following algorithm is used [6]:

1. Eliminate consonant modifiers from the word by finding the middle row using bounding box and also calculate horizontal profile to recognize the bottom base line.
2. Remove vowel modifiers by finding the top base line.
3. Using the vertical profile divide the base characters using the white space between them. Then include vowel and consonant modifiers using nearest neighborhood method with horizontal correlation heuristics.

The advantage and disadvantage of each segmentation algorithm [9] is given below in the table 5.1

Table 5.1: Advantage and disadvantage of segmentation algorithm

Segmentation methods	Advantage	Disadvantage
Connected Component based	The Connected Components technique solves the overlapping character segmentation problem	It separates simple characters into their basic glyphs.
Projection profile method	This method is appropriate for segmenting image documents that are well spaced without overlapping and touching	When the characters are overlapped or touched this method can't segment
Segmentation methods	Advantage	Disadvantage
Particle swarm optimization	This method is right for segmenting image documents contained overlapping characters documents	Tamil scripts are consists of two parts, namely the basic character and a modifier symbol corresponding to each of the basic character. If the distance between the basic character and the modifier symbol is more, this method couldn't fragment it correctly.

VI. FEATURE EXTRACTION

The fundamental components of Tamil characters, such as loops, straight lines, zigzag lines and the position of the connection between the loop and the straight line are extracted by this module. In general a Tamil character consist three loops, three zigzag lines and five straight lines. After, this module extracts the basic components from the Tamil character image. There are two main basic components that the system extracts, which are: 1) a stroke extraction, and 2) a loop extraction. Both components include the following things

A. Stroke extraction

It is a sub-process that extracts the type of a line by sorting out a character into 5 X 5 pixels image. It is a sub-process that extracts the type of a line by sorting out a character into 5 X 5 pixels image blocks and finding the slope of a line in each block. Finally, it analyzed and classified a line type of a character in four categories, which are vertical line, horizontal line, zigzag line, and the tail line. Each line category has the following information.

1. Vertical Stroke Analysis

This method that extracts the vertical stroke lines in the Tamil character. The system divides one character into three vertical regions such as Left Region, Middle Region, and Right Region.

2. **Horizontal Stroke Analysis**

Horizontal stroke line in the Tamil character is extracted by using horizontal stroke Analysis. The system separates a character into three regions, which are Upper Region, Central Region, and Lower Region.

3. Zigzag Stroke Analysis

The zigzag stroke analysis in system means the stroke that contains a turning point. The zigzag line in a Tamil character is divided into 3 zones, namely Top Zone, Middle Left Zone, and Bottom Zone.

4. Tail Stroke Analysis

The tail stroke of the Tamil character is extracted by using Tail Stroke analysis function. Some Tamil characters have a long tail stroke. The system focuses on the location of a tail. There are 2 positions, which are Upper tail and Lower tail.

B. Loop Detection

It is a sub-process that extracts three characteristics of the loop, which are number of loops in a character image, position of each loop in a character image, and type of the loop.

VII. CLASSIFICATION

A. Feed Forward Backpropogation Neural Network

To execute the character recognition process, each character segmented from the manuscript given as input. In this classification technique, q number of FFBNN (q-FFBNN) is utilized to complete the recognition process [7]. In q-FFBNN, there is H_d number of hidden layers and one output layer, which indicate that the resultant input character is recognized or unrecognized. The q-FFBNN is well trained by this segmented characters and provided an exact result for the corresponding input. The q-FFBNN contains q input units (the character q is recognized in q-FFBNN and other character are unrecognized), one output units and H_d hidden units. Initially, the input value is forwarded to the hidden layer and then to the output layer. Hence this process is known as forward pass of the back propagation algorithm. Every node in the hidden layer gets input from the input layer, which are multiplexed with appropriate weights and summed. The hidden layer input value calculation function is called as bias function, which is described below:

$$r_{a_{ij}}^{z} = \alpha + \sum_{h=1}^{\mathrm{h_d}} \left(w_h \, \mathrm{c_q} \, \mathrm{a_{ij}^{z}} \, \mathrm{h} \right) \qquad (1)$$

In Equation (1), is the input segmented character from the document. The output of the hidden node is the non-linear transformation of the resulting sum. Same process is followed in the output layer. The following Equation (2) denotes the starting function of the output layer. The output values from the output layer square measured compared

with target values and also the learning error rate for the neural network is computed, which is given in Equation (3):

$$A = \frac{1}{1+ e^{-}r_{aij}^{z}} \quad (2)$$

$$\delta = \frac{1}{H_d} \sum_{h=1}^{H_d} D_h^{a_{ij}^z} - A_h^{a_{ij}^z} \quad (3)$$

In Equation (3), is the learning error rate of the q- FFBNN, In Equation (3), d is the learning error rate of the q-FFBNN. The error between the nodes is transmitted back to the hidden layer and this process is called the backward pass of the back propagation algorithm. The reduction of error by back propagation algorithm is described in the subsequent steps.

Initially, the weights are assigned to hidden layer neurons. The input layer has a constant weight, whereas the weights for output layer neurons are chosen arbitrarily. Subsequently, the bias function and output layer activation function are computed by using the Equation (1 and 2).

Next, the back propagation error is computed for each node and the weights are updated by using the Equation (4):

$$w_{a_{ij}}^{z} \text{ h} = w_{a_{ij}}^{z} + \Delta w_{a_{ij}}^{z} \text{ h} \quad (4)$$

Where the weight is changed, which is given as Equation (5):

$$\Delta w_{a_{ij}}^{z} \text{ h} = \delta . r_{a_{ij}}^{z} .\text{h.E}(\phi) \quad (5)$$

Where, δ is the learning rate that normally ranges from 0.2 to 0.5 and $E(\phi)$ is the BP error.

The bias function, activation function and BP error calculation process are continued till the BP error gets reduced i.e., $E(\phi) < 0.1$. If the BP error reaches a minimum value, then the q-FFBNN is well trained by the segmented characters for performing the character recognition the well trained q.

During the testing, the unknown Tamil palm leaf manuscripts images to be taken for analyzing the performance of the trained FFBNN. The unknown palm leaf images are given to the preprocessing, line and character segmentation process which is already given in section 5.1. The segmented characters are given to the well trained q-FFBNN to check whether the given input characters are recognized or unrecognized.

B. Genetic Algorithm

A GAs is an optimization and search method to find better solutions [8]. The evolution starts from a population and it is completely random. Individuals occur in generations. In each generation, the fitness of the entire population is evaluated, and many individuals are chosen from the present population based on their fitness. These are adapted mutated, or again combine to create a new population, which becomes present in the next iteration of the algorithm. Generally the solutions are represented in strings of 0s and 1s, while different encodings are also achievable.

A GAs is an optimization and search method to find better solutions. The evolution starts from a population and it is completely random. Individuals occur in generations. In each generation, the fitness of the entire population is evaluated, and many individuals are chosen from the present population based on their fitness. These are adapted mutated, or again combine to create a new population, which becomes present in the next iteration of the algorithm. Generally the solutions are represented in strings of 0s and 1s, while different encodings are also achievable. , the extracted character details are kept in the form of bits in a genetic algorithm. Finally, the system displays the best fitness chromosome for the recognition result. When the characteristics of the characters in the sub-word are identified, the next phase is to recognize the characters of the sub-word; while the features of characters in the sub word are identified .The genetic algorithm method will be used for this purpose. First system retrieves the image from the database then it divides that image into lines. After lines are segmented into words and then segment words into sub-words. Now we classify the image and after resolve the number of peaks in that image. Then we detect stroke, loop, location of loop, stroke connection in the peak and after that determine the complementary character. Now we compute the height and width of the peak and determine left and right connection. After it send this peak's string to the genetic algorithm.

In genetic algorithm we have initial population and then we applying three operators in which first is selection which selects the strings and then we apply crossover operator which recombining those selected strings and after that we apply mutation operator those changes strings of 0's and 1's form. The chromosome bit string from the chromosome generation function is used to identify a character by comparing the fitness value of an unknown character with all Tamil characters in the database. By combining all features such as

- 3-bit for number of loops
- 27-bit for location of every loop
- 24-bit for loop connected with a straight line
- 12-bit for location of the lines are extracted from previous module and finally chromosome function produce the Tamil character chromosome.There are 66-bit chromosomes in the Tamil character.

Now we apply condition and check optimization criteria met or not, if met then selects the best string which is our solution and if not then send it to in the initial population. The highest fitness value is the recognition outcome. The fitness value is computed by using Equation 1 as the following:

$$fitness\ value = \sum_{i=1}^{66} |(\text{S}_\text{i} + 1.0) - (\text{L}_\text{i} + 1.0)| * w_i$$

Where S is a chromosome bit string in database Y is a Chromosome bit string of an unknown character W is weight of each chromosome bit string.

VIII. CONCLUSION AND FUTURE WORK

This survey provides an overview of the current research in Tamil Palm manuscript Recognition systems. The performance of various preprocessing, segmentation, and feature extraction and classification methods has also been analyzed, which gives a picture of efficient factor of each method. However, there is no standard solution to identify all Tamil characters with reasonable accuracy. Various methods have been used in each phase of the recognition process, whereas each approach provides solution only for few character sets. Challenges still prevails in the recognition of normal as well as abnormal writing, slanting characters, similar shaped characters, joined characters, curves and so on during recognition process. These are only some basic issues which can be overcome through future extension of character recognition.

REFERENCES

[1] Vikas J Dongre and Vijay H Manka, "Devanagari Document Segmentation Using Histogram Approach", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, pp. 46 -53, August 2011

[2] C V Lakshmi, C Patardhan "A Multi-font OCR System for printed Telugu Text.", Proceeding of LEC'02, IEEE, 2002.

[3] L. Likforman-Sulem, A. Zahour, B. Taconet,"Text line segmentation of historical documents: a survey", International journal of Document Analysis and Recognition,Vol 9, 2007, pp. 123 – 138.

[4] Itay Bar-Yosef et, al, "Line segmentation for degraded handwritten historical documents".

[5] R.Sanjeev Kunte and R D Sudhaker Samuel," A Simple and efficient optical character recognition system for basic symbols in printed kannada text", Sadhana, Vol 32, Part 5, pp. 521 – 533,2007.

[6] M Swamy Das et. al, "Segmentation of Overlapping Text Lines, Characters in Printed Telugu Text Document Images", International Journal of Engineering Science and Technology, Vol. 2, No.11,pp. 6606 – 6610,2010.

[7] Ramya J. and B. Parvathavarthini, "Feed Forward Back Propagation Neural Network Base Character Recognition System for Tamil Palm Leaf Manuscripts", Journal of Computer Science Volume 10, Issue 4 Pages 660-670, 2014.

[8] E. K. Vellingiriraj, P. Balasubramanie, "Recognition of Ancient Tamil Handwritten Characters in Palm Manuscripts Using Genetic Algorithm", International Journal of Scientific Engineering and Technology, Volume 2 Issue 5, pp: 342-346, 2013.

[9] N. Sridevi, P.Subashini,"Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques", International Journal of Computer Applications, Volume 52– No.14, 2012.

[10] E.K.Vellingiriraj, Dr. P.Balasubramanie, "Recognition of Ancient Tamil Handwritten Characters in Historical Documents by Boolean Matrix and BFS Graph", International Journal of Computer Science and Technology, Vol. 5, SPL - 1, 2014.

[11] R.C. Gonzalez and R.E. Woods. (2004): Digital Image Processing, Pearson Education.

[12] Stephen Marchand Maillet ,"Binary Digital Image Processing- A Discrete Approach", 1999.

[13] Oliveira .S.L., S. A. Britto, and R. Sabourin, " Optimizing Class-Related Thresholds with Particle Swarm Optimization", Proceeding of International Joint Conference on Neural Networks, IEEE, Montreal, Canada, July 31 – August 4, 2005,pp. 1511 – 1516.