Recognition of Ancient Tamil Handwritten Characters in Historical Documents by Boolean Matrix and BFS Graph

[1]E.K.Vellingiriraj, [2]Dr. P.Balasubramanie
[1,2]Dept. of CSE, Kongu Engineering College, Perundurai, Erode, Tamilnadu, India

Abstract
Tamil is one of the oldest languages in the world with rich literature. In the ancient days, the writers, especially in Tamilnadu, used palm leaves to encrypt their writing. A very good example of the usage of Palm leaf manuscripts to store the history is Tamil grammar book named Tolkappiyam which was written during 4th B.C. The ancient literature includes many palm leaf manuscripts that contain Sangam works, classics, Saiva, Vaishnava and Jain works, medical works, food, astronomy & astrology, vaastu & Kaama shastra, jewellery, music, dance & drama, medicine, Siddha and so on. Over the 3,500 Tamil manuscripts are available in Saraswathi Mahal Library located in Thanjavur, Taminadu, India. In this library, only a few palm leaf manuscripts are digitalized and many are to be digitalized so as to enable quick reference in the future. The objective of the proposed research is to develop the system that can recognize Tamil characters from palm manuscripts and convert them into text format using the Breadth First Search (BFS).

Keywords
Ancient Tamil Handwritten Characters, Booloean Matrix, BFS Graph, Historical Documents

I. Introduction

Vattezhutthu alphabet (means rounded letters) is an abugida writing system originating from the ancient Tamil people of Southern India. Developed from the Tamili (Tamil-Brahmi), Vatteluttu is one of the three main alphabet systems developed by Tamil people to write the Proto-Tamil language, alongside the more modern Grantha alphabet (Pallava or Grantha Tamil) and Tamil alphabet. The syllabic alphabet is attested from the 6th century CE to the 14th century in present day Tamil Nadu. It was also an ancient writing system used for writing the Tamil language after the 2nd century BCE replacing an older Tamil-Brahmi script based on the Brahmi writing system. Inscriptional records in the Tamil language date from 300 BCE to 1800 and have undergone varying changes through history. Handwritten character recognition is one of the most difficult tasks in the pattern recognition system. There are a lot of difficult things that need many image processing techniques to solve, for examples: 1) how to separate cursive characters into an individual character, 2) how to recognize unlimited character fonts and written styles, and 3) how to distinguish characters that have the same shape but different meaning such as the character o and number 0. Many researchers try to apply many techniques for breaking through the complex problems of handwritten character recognition.

II. Literature Reviews

Historically, handwritten character recognition applications used three major approaches; the statistical approach, the structural or syntactic approach, and the neural network-based approach. This section reviews handwritten character recognition applications based on these three approaches.

A. Statistical Analysis Approach

Statistical Pattern Recognition uses statistical and/or probabilities 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 to express pattern classification in terms of structural information. [2-8].

B. Structural or Syntactic Analysis Approach

Syntactic Pattern Recognition uses syntactic or structural information of patterns to generate knowledge that is related to patterns. This approach extracts the similarity of patterns and builds pattern syntax or structural rules. The information of pattern syntax rules is used to explain, classify and recognize unknown patterns. This approach is uitable for building a handwritten character recognition system because it uses astructural approach to build unlimited handwritten character patterns syntax. A limitation of this approach is the difficulty to build learning structural rules. [9-13].

C. Neural Network Based Approach

Neural Pattern Recognition emulates knowledge of how a biological neural system stores and manipulates information. This artificial neural system is called “neural networks”. The notion is that an artificial neural network can solve all problems in automatic reasoning, including a pattern recognition problem. This approach classifies patterns through predictable properties of neural networks. A limitation of this approach is a little amount of semantic information from a network. [14-16].

III. Methodology

Here we present all details of our system design. First, we start the overall framework of the ancient Tamil handwritten character recognition system. Then we give the components details.

A. System Architecture Overview

Fig. 1: Framework of the Ancient Tamil Handwritten Character Recognition in Palm Manuscripts

B. System Structure Chart

Based on the system framework in the previous section, we convert the Tamil palm leaf image to Tamil text format. This framework includes, i) Image scanning ii) Image preprocessing iii) Feature extraction iv) Character recognition v) Convert the Text.

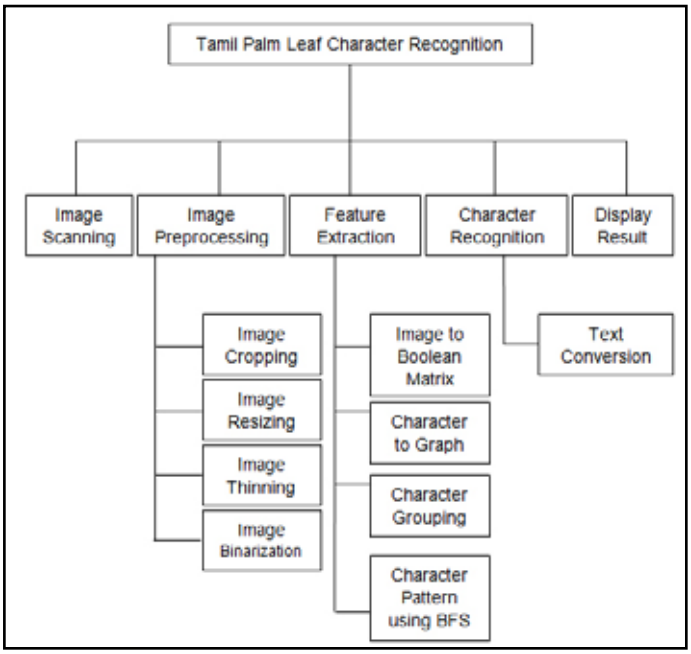


Fig. 2: The Structure Chart of Ancient Tamil Hand Written Characters Recognition by Image Zoning Using the Boolean Matrix

1. Image Scanning

In the first stage, the Tamil palm leaf manuscript collect from the various places from different centuries. The palm manuscript scans by 4800 dpi scanner and store the Jpeg image.

2. Image Preprocessing

In the image preprocessing module, the system prepares a palm manuscript handwritten character image for the feature extraction module. The image preprocessing stage consists of for sub-process. That is (i) image cropping (ii) Segmentation (iii) image resizing d) image thicken and e) Image binarization. Each of these sub-process details below:

(i). Image Cropping

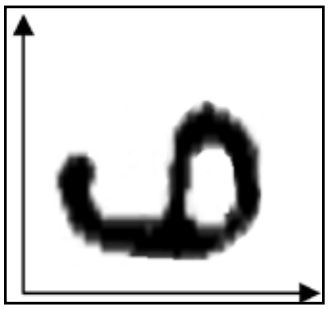
The palm leaf image from the scanning stage has the white space. In the palm image the characters are easily to identify from the each pixel color. The characters are darken color in the palm script remaining spaces are brown color.

(ii). Segmentation

Using the graphemes extraction [17] technique or edge detection method, each word can be segmented by the characters.

(iii). Image Resizing

After cropping the particular character, it may be different size of each character. So we need to change each character is same size. The character image to resized 100X100 pixels image.



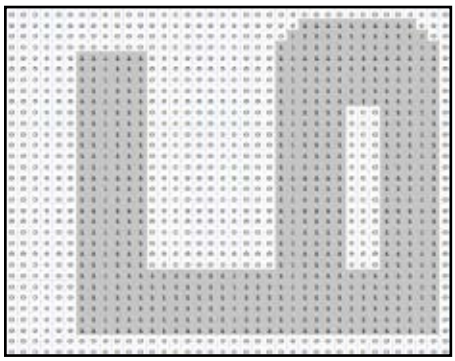
100 X 100 Pixels

(iv). Image Thinning

The darken pixel (i.e. a character) is convert the thinning character. A thicken character easily to extract the thin character by using a nearest pixel darken to lighten color change to particular range.

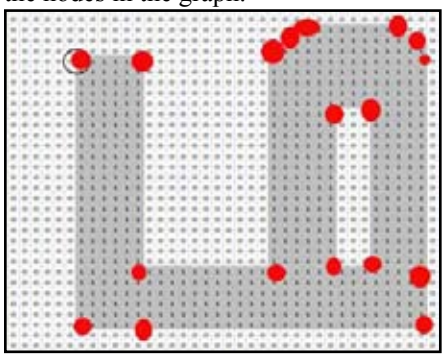
(v). Image Binarization

A character is store the Boolean matrix to store either 0's or 1's. Using the Image Zoning technique[5], the dark pixel is stored the 1's and lighten pixel is stored the 0's.



(vi). Boolean Matrix to Graph

In the above Boolean matrix find the 1's in each row and each column. Point out the left, right, top or bottom matrix if the 0's are available. If it is changed that rows and columns are point out the nodes in the graph.



(vi). Graph Generate

Each point noted in the above picture then draw the graph G=<V,E> edges and vertex in graph G.

(vii). Using the Breadth

First Search algorithm each node is travel to the nearest node through the edges.

(viii). Pattern Matching

This travels is compare to all the existing avail ancient tamil characters in the data set, if the match is found then convert the equivalent Tamil text is changed.

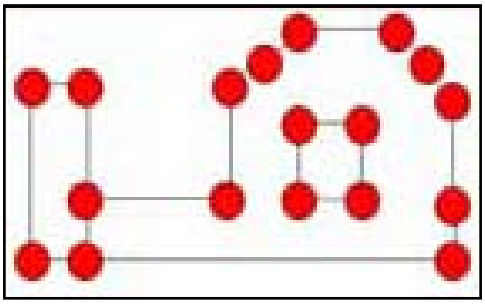


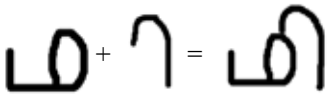
Fig. 5: Each Individual Script Is Stored in Boolean Matrix

(ix). Image Grouping

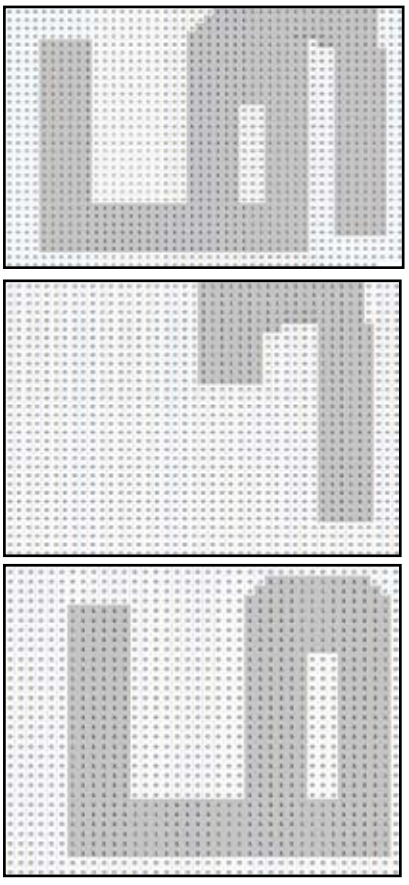
Palm manuscripts were used in different centuries and the words appear in them are of different styles and strokes (Fig 5). Each individual script is stored in Boolean matrix.

IV. Character Modeling

A character model is a record of all the characters set that are of equal Boolean matrix. The combination of the two Boolean matrix is also equal to the single character. For example:



The equivalent Boolean matrix is given below:



V. Conclusion

In this paper, we have proposed a simple method for converting ancient Tamil handwritten scripts into text format. There are thousands of Tamil palm manuscripts that are yet to be digitalized. The aim of this paper is to convert the palm manuscript image into digitized text format. However, our method has some difficulties in handling cases such as cursive Tamil script, merging of two Boolean matrixes, and a hole in palm manuscript image. These are only some basic issues which can be overcome through future extension of character recognition.

References

[1] Hyung Il Koo, Nam Ik Cho,"Text Line Extraction Chinese Documents Based on an Energy Minimization Framework", IEEE Trans. On Image Processing, Vol. 21, No. 3, pp. 1169-1175, Mar 2012.

[2] G. Nagy, S. Seth, M. Viswanathan,"A Prototype Document Image Analysis System for Technical Journals," Computer, Vol. 25, No. 7, pp. 10–22, Jul. 1992.

[3] F. Shafait, D. Keysers, T. M. Breuel, "Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 6, pp. 941–954, Jun. 2008.

[4] Y.Liang, M.C.Fairhurst, R.M.Guest, "A Synthesisd Word Approach to Word Retrieval in Handwritten Documents", Elsevier Pattern Recognition, Vol.45, PP 4225-4236, June 2012.

[5] Giuseppe Pirlo, Donato Impedovo, "Adaptive Membership Functions for Handwritten Character Recognition by Voronoi-Based Image Zoning", IEEE Trans on Image Processing, Vol 21, No. 9, pp. 3827-3836, Sep 2012.

[6] Chomtip Pornpanomchai, Verachag Wongsawangtham, Satheanpong Jeungudomporn, and Nannaphat Chatsumpun, "Thai Handwritten Character Recognition by Genetic Algorithm (THCRGA)", IACSIT Journal of Engineering and Technology, Vol. 3, No 2, Apr 2011.

[7] Qiu-Fend Wang, Fei Yin, Cheng-Lin Liu,"Handwritten Chinese Text Recognition by Integrating Multiple Contexts, IEEE Trans on Pattern Analysis and Machine Intelligence,

Vol. 34, No. 8, Aug 2012.
[8] Tiji M Jose, Amitabh Wahi,"Recognition of Tamil Handwritten Characters using Daubechies Wavelet Transforms and Feed-forward Back Propagation Network", IJCA, Vol. 64, No. 8, pp. 0975-8887, Feb 2013.
[9] Jin Chen, Daniel Lopresti," Model Based Ruling Line Detection in Noisy Handwritten Documents", Pattern Recognition Letters, Elsevier, 2012.
[10] A Bharath, Sriganesh Madhvanath,"HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts", IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. 34, No. 4, Apr 2012.
[11] Chomtip Pornpanomchai, Dentcho N. Batanov, Nicholas Dimmitt,"Recognizing Thai handwritten characters and words for humancomputer interaction", International Journal of Human-Computer Studies, pp. 259-279, 2001.
[12] Chomtip Pornpanomchai, Pattara Panyasrivarom, Nuttakit Pisitviroj, Piyaphume Prutkraiwat,"Thai Handwritten Character Recognition by Euclidean Distance", The 2nd International Conference on Digital Image Processing (ICDIP 2010), pp. 53-58, 2010.
[13] Chomtip Pornpanomchai, Montri Daveloh,"Printed Thai Character Recognition by Genetic Algorithm", The International Conference on Machine Learning and Cybernetics, pp. 3354-3359, 2007.
[14] Boontee Kruatrachue, Nattachat Pantrakarn, Kritawan Siriboon,"State Machine Induction with Positive and Negative for Thai Character Recognition", The International Conference on Communications, Circuits and Systems, pp. 971-975, 2007.
[15] Parinya Sanguansat, Widhyakorn Asdornwised and Somchai Jitapunkul, "Online Thai Handwritten Character Recognition Using Hidden Markov Models and Support Vector Machines", ThInternational Symposium on Communications and Information Technologies, pp. 492-497, 2004.
[16] Rud Budsayaplakorn, Widhayakorn Asdornwised and Somchai Jitapunkul,"On-line Thai handwritten character recognition using hidden Markov model and fuzzy logic", The IEEE 13th Workshop on Neural Networks for Signal Processing, pp. 537-546, 2003.
[17] Kritawan Siriboon, Apirak Jirayusakul and Boontee Kruatrachue,"HMM topology selection for on-line Thai handwriting recognition", The First International Symposium on Cyber Worlds, pp. 142-145, 2002.
[18] Arrak Pornchaikajornsak and Arit Thammano, "Handwritten Thai character recognition using fuzzy membership function and fuzzy ARTMAP", The IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp. 40-44, 2003.
[19] Supachai Tangwongsan, Orawan Jungthanawong, "Refinement of stroke structure for printed Thai character recognition", The 9th International Conference on Signal Processing, pp. 1504-1507, 2008.
[20] Khampheth Bounnady, Boontee Kruatrachue and Takenobu Matsuura,"Online Unconstrained Handwritten Thai Character Recognition Using Multiple Representations", The International Symposium on Communications and Information Technologies, pp. 135-140, 2008.
[21] Jarernsri L. Mitrpanont, Urairat Limkonglap,"Using Contour Analysis to Improve Feature Extraction in Thai Handwritten Character Recognition Systems", The 7th IEEE International Conference on Computer and Information Technology, pp. 668-673, 2007.
[22] Hyung Il Koo, Nam Ik Cho,"Text Line Extraction Chinese Documents Based on an Energy Minimization Framework", IEEE Trans. On Image Processing, Vol. 21, No. 3, pp. 1169-1175, Mar 2012.
[23] G. Nagy, S. Seth, M. Viswanathan,"A Prototype Document Image Analysis System for Technical Journals", Computer, Vol. 25, No. 7, pp. 10–22, Jul. 1992.

Dr.P.Balasubramanie received his M.Sc. degree in Maths from Bharathiar University, Coimbatore, India in 1989, the M.Tech. degree received from Archarya Nagarjuna University, AP, India, and the Ph.D. degree in Theoretical of Computer Science from Anna University, Chennai, India in 1996. He is working as Professor, Department of Computer Science & Engineering in Kongu Engineering College, Perudurai, Erode, India. His research interests include Fuzzy Logic, Networks, Image Processing and Natural Language Processing.

E.K.Vellingiriraj received his M.C.A. degree from Periyar Unversity, Salem, India in 2002 and Completed M.E. in Software Engineering from Anna University, Coimbatore, India in 2010. His Self Interested completed M.A. Tamil in TNOU, Chennai. Now he is doing research in Anna University, Chennai in Natural Language Processing of Character Recognition from Historical Palm Manuscripts.