IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 5, March 2019 E-ISSN: 2347-2693

Contour based Character Segmentation and Nguyen-Widrow Weight Generation for Classification of Tamil Palm Leaf Script Characters - Machine Learning Approach

Poornima Devi. M[1], M. Sornam[2*]

[1,2]Department of Computer Science, University of Madras, Chennai, India

[*]*Corresponding Author: madasamy.sornam@gmail.com, Tel.: +91-044-2220 2904*

DOI: https://doi.org/10.26438/ijcse/v7si5.118123 | Available online at: www.ijcseonline.org

Abstract—The main aim of this work is the classification of Tamil palm leaf manuscript segmented characters using Machine Learning approach. For the segmentation of characters, first the images of the palm leaf manuscripts were allowed to preprocessing which includes filtering and thresholding. After the preprocessing stage, the preprocessed images were allowed for character segmentation using contour based bounding box segmentation. Then the segmented Tamil palm leaf manuscript characters were labelled with different classes for classification. To classify the characters Adaptive Backpropagation Neural Network (ABPN) with Shannon activation function was used with Nguyen Widrow weight initialization. For neural network, normally we use random initialization to generate the weights. Rather than random initialization here Nguyen-Widrow weight initialization technique was implemented. For comparison ABPN with Shannon activation function (method 1) and ABPN with Shannon activation function using Nguyen-Widrow initialization was used, from this ABPN with Shannon activation function using Nguyen-Widrow gives 96% of accuracy for Tamil palm leaf character classification.

Keywords—ABPN, Bounding box, Convex hull, Contour, Shannon, Machine Learning.

I. INTRODUCTION

The main intension of this work is to classify the Tamil handwritten palm leaf manuscript character using Machine Learning approach. Tamil language is considered as an official language in the state Tamil Nadu and also a classical language of the country India. It is also widely spoked by many states in India includes Kerala, Karnataka etc. The countries which announced Tamil as an official language is Sri Lanka and Singapore. It was widely spoken by the peoples in the countries like South Africa, Malaysia and Mauritius. Tamil language is considered as most challenging language because of the complexity of the characters and century based characters. It characters contains dots, curves, circles, semi circles, lines etc. So the handwritten character recognition is more challenging task especially for Tamil palm leaf manuscript. Tamil language contains totally 247 characters except Grantham characters [12] as shown in Figure 1.

Tamil language contains basic 30 characters which include 12 characters of soul-letters, 18 characters of body-letters and one special character. These are also known as Uyirezhuthu, Meyezhuthu and Ayudhaezhuthu. It also contains 216 characters of soul-body-letters which is also known as Uyiremeyezhuthu. Meyezhuthu can be further classified as vallinam, mellinam and idayinam which can be pronounced based on the sound kuril and nedil.

Figure 1.6 Granta characters.

Tamil palm leaf manuscript is the one of the oldest method of writing to save the information to future generation. It was written in dried palm leaf and it is considered as one of the handwritten script. Handwritten character recognition is most challengeable task because of the writing style of the individual will differ and the writing style will depend on the person's state of mind. Most of the palm leaves are related to medical science, foods, Ayurveda, astrology, astronomy etc. These palm leaf are 200 to 300 years old due to lack of maintenance its get affected by insects or by natural disasters as shown in Figure 2.

Figure 2. Sample of Pam leaf manuscript

This paper is organized as follows, Section I contain the introduction, Section II contain the related work, Section III contain proposed work for Tamil palm leaf classification, Section IV explain the experimental result and discussionand section V contain conclusion.

II. RELATED WORK

NarahariSastryPanyam et al.[8] developed a method to identify the Telugu palm leaf manuscripts by using Optical Character recognition with two dimensional Discrete wavelet transform, two dimensional discrete cosine transform and two dimensional fast Fourier transform as a feature extraction technique and for classification K-Nearest Neighbor classifier was used which produced 96.4% of accuracy.

Karthigaiselvi et al. [5] proposed a novel method for printed Tamil characters feature extraction. Here structural feature extraction method is implemented to extract vertical and horizontal projection of the characters. It includes lower zone, middle zone and upper zone. These extracted features were allowed as an input to neural network and achieved 99.67% of accuracy for printed Tamil characters.

Kiruba et al. [7] implemented the work to segment the characters from the Tamil palm leaf manuscript by eliminating the background from an input image. The line segmentation and character segmentation were done using histogram based segmentation.

Vijaya Lakshmi [14] proposed the method to recognize the palm leaf characters from Telugu manuscript using 3D based feature extraction technique. This method produced 94.68% of accuracy with K-Nearest Neighbor classifier.

Vellingiriraj [13] developed the method to identify the Tamil palm leaf script and convert it to text document by using Breath First Search Algorithm. This methodology includes image scanning, preprocessing to remove noise, extracting features and recognizing the characters.

Yaping Zhang et al. [15] implemented the method for character recognition with 98.29% of accuracy for handwritten Chinese character using Deep Convolutional Neural Network. This work is useful to know the knowledge of the printed document data with novel adversarial feature learning model.

KavithaSubramani [6] developed the approach for the improvement of degraded palm leaf script quality. It includes preprocessing, grayscale conversion, resize, denoizing. For binarization Otsu method and mean shifting algorithm was implemented and for post processing trimmed mean filter was implemented.

Yi-Chao Wu et al. [16] achieved 95.88% of accuracy for the recognition of Chinese characters using Convolutional Neural Network. Feedforward Neural Network Language Model (FNN LM) and Recurrent Neural Network Language Model (RNN LM) are the two types of characters level neural network language model used to form the hybrid method called Back-off N-gram.

Nibaran Das et al. [9] used the approach to recognize the OCR handwritten numerals using the classifier Support Vector Machine (SVM). Proposed the Quad Tree based hierarchically derived Longest-Run (QTLR) with Principal Component Analysis (PCA) and QTLR with Modular PCA (MPCA); from these two methods MPCA with QTLR achieved better performance than the PCA with QTLR.

Amit Choudhary et al. [1] achieved 85.62% of accuracy for the classification of English handwritten character using Optical Character Recognition (OCR) system. To achieve this accuracy Multilayer feedforward backpropagation neural network algorithm was used for classification.

III. PROPOSED WORK

The main objective of the work is to make Tamil palm leaf manuscript to readable format. So that everyone can read and understand the facts behind the hidden content of manuscript. The digitalization is required in the field of Tamil palm leaf manuscript to preserve the historical details and to know the hidden facts. To perform this, first the dataset of Palm leaf images were collected from Tamil Nadu Archaeological Department then preprocessed which includes converting to grayscale images, then the grayscaled images were allowed to filtering technique to remove noise from the images. Here median filter is suited from palm leaf scripts. The filtered images were then allowed for binarization using Adaptive Mean thresholding method. Then the output of the threshold image was allowed to segment the characters using contours based bounding box segmentation. The segmented characters are then resized to make all the characters images in same dimension. The resized characters were saved as different class labels for character dataset [11].

The resized class labelled characters was used for classification using Adaptive Backpropagation Neural

Network (ABPN). The proposed method contains two activation functions: Shannon activation function and sigmoid activation function. The resized segmented characters were in the 8x8 dimensions. For training the network, intensity of pixels of an image was considered as an input for single characters. So there are 64 input nodes in the input layer, 55 nodes in the hidden layer and 5 nodes in the output layer. For this work 18 characters class labels are considered and the output was represented by 2^5. For weight generation normally random initialization will be used, rather than using random initialization here Nguyen Widrow weight generation was implemented to optimize the weights. This proposed methodology achieved 96% of accuracy for classification of Tamil palm leaf manuscript characters.The architecture of the proposed work has been shown in Figure. 3.

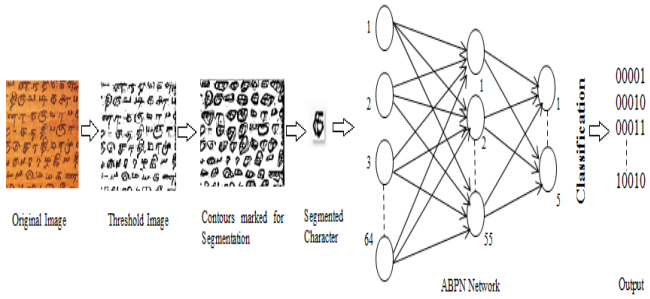

Figure 3.Architecture of the proposed work.

A. Conotur based Convex Hull Bounding Box Segmentation

Here the segmentation of character was done without performing line segmentation and word segmentation by using contour based convex hull bounding box segmentation. For this first the region was spotted from the input image and convex hull points were noticed. Then the polygon lines were drawn with in the observed region where the convex hull points were marked. At last bounding box were drawn for each character detected region as shown in Figure 4 and Figure 5. The main disadvantage of this segmentation is that the connected characters cannot be segmented correctly.

Here contour based segmentation was used rather than edge detection algorithm. The important differentiation between contour detection [12] and edge detection is that the edge detection algorithm will detect the edges from an images but it will not detect the curves in an images. Whereas Contours detects the curves which is more useful for character segmentation and it will also detect the connected edges. The contour based convex hull bounding box segmentation is well suited for Tamil palm leaf character segmentation.

Figure4. Points marked to detect the characters using contour based convex hull bounding box segmentation.

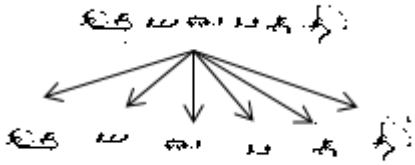

Figure5.Segmented characters from Tamil palm leaf manuscript.

B. Adaptive Backpropagation Neural Network (ABPN) with Shannon activation function using Nguyen-Widrow weight generation technique

Adaptive Backpropagation Neural Network (ABPN) is a modified BPN in the error patterns which as linear and non-linear error patterns in output layer. In ABPN, global adaption technique was used to adapt the weight parameters [4,10]. For this proposed work, two activation functions were implemented; sigmoid activation function was used from input to hidden layer as shown in equation (1) and Shannon activation function was used from hidden to output layer as shown in equation (2). The derivative of Shannon activation function was shown in equation (3).

$$f(\delta_{hid}) = \frac{1}{1+ e^{-\delta_{hid}}} \quad (1)$$

$$f(\delta_x) = \frac{\sin(1.75x)-\sin(0.75x)}{1.75\ x} \quad (2)$$

$$f'(\delta_{x_i}) = \frac{((1.3125*\ x*\cos(1.75*x))-0.5625*x*\cos(0.75*x)))}{-((0.75*\sin(1.75*x))-(0.75*\sin(0.75*x)))}{(0.5625\ x^2)} \quad (3)$$

For weight generation Nguyen-Widrow weight generation method was implemented to optimize the weights from input layer to hidden layer and from hidden layer to output layer. Nguyen-Widrow weight generation from input layer to hidden layer was implemented using the equation from (4) to (6).

$$\beta_i = 0.7\ \text{hid}^{\frac{1}{n_i}} \quad (4)$$

$$k_i = \sqrt{\sum W_{ij}^2} \quad (5)$$

$$W_{ij_{t+1}} = \frac{\beta W_{ij}}{k_i} \quad (6)$$

Nguyen-Widrow weight generation from hidden layer to output layer was implemented using the equation from (7) to (9)

$$\beta_i = 0.7\ o^{\frac{1}{hid_i}} \quad (7)$$

$$k_i = \sqrt{\sum W_{ij}^2} \quad (8)$$

$$W_{ij_{t+1}} = \frac{\beta W_{ij}}{k_i} \quad (9)$$

where
n_i – number of input neurons.
hid– number of hidden neurons.
o – number of output neurons.
W_{ij}– weights initialized with random values.
$W_{ij_{t+1}}$- equals to adjusted weights.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Tamil palm leaf Manuscript datasetwere collected from Tamil Nadu Archaeological department, Chennai, India. There are huge number of palm leaf manuscript are available in the fields like medical science, astrological, science, historical etc. The dataset used for this work is astrological palm leaf manuscript especially for lagnams written by SubtharishiNaash[11]; basically lagnams can be classified as 11 types and each lagnam leaf will have different parts.

Different types of lagnams and their parts are;mesa lagnamwith 40 parts, virusabha lagnam with 26 parts,mithuna lagnam with 22 parts, kataka lagnam with 22 parts, simha lagnam with 18 parts, thula lagnam with 23 parts, virchika lagnam with 28 parts, thanoor lagnam with 15 parts, magara lagnam with 13 parts, gumpa lagnam with 25 parts and meena lagnam with 14 parts. Each and every part in lagnam leaf contains 10 leaves, totally 2460 palm leaf images with 198 x 2063 dimensions [11].

The image dataset and Neural Network was implemented using Python 3.7 in Anaconda Navigator (Spyder) with 4 GB RAM with Intel Pentium Processor CPU 2127U @ 1.90GHz of 64 bit Windows Operation System.

After segmentation each characters segmented from Tamil palm leaf manuscript are labelled to different classes. For this proposed work 18 classes were used as shown in Figure 6.

Characters	Class Labels	Characters	Class Lables	Characters	Class Labels
க	1	த	7	ல	13
ங	2	ந	8	வ	14
ச	3	ப	9	ழ	15
ஞ	4	ம	10	ள	16
ட	5	ய	11	ற	17
ண	6	ர	12	ன	18

Figure 6. Characters with corresponding class labels.

To visually identify the correct classification and misclassification of characters confusion matrix was drawn. The Performance metrics used are; precision, recall, f1-score, classification accuracy and classification error.

From the confusion matrix for ABPN with Shannon activation function using Nguyen-Widrow weight initialization shown in the Figure 7, has totally 18 character class labels which were started from 0 to 17. It was clearly noted that class label 0 has totally 15 characters where 1 character is misclassified as class label 6, then class label 7 has totally only one character which is misclassified as class label 1, then class label 8 has totally 24 characters where 1

Table 1: Comparison: ABPN with Shannon activation function and Nguyen-Widrow weight initialization.

S.No	Network	Network Architecture	Epochs	Error	Average Precision	Average Recall	Average F1-score	Classification Accuracy	Classification Error
1	ABPN with Shannon uisng Nguyen-Widrow (Proposed)	64-55-5	190	0.0003	0.96	0.95	0.95	96%	4%
2	ABPN with Shannon	64-55-5	410	0.00036	0.96	0.93	0.94	94%	6%

character is misclassified as class label 16, class label 9 has totally 24 characters where 1 character is misclassified as class label 8, class label 17 has totally 7 characters where 1 character is misclassified as class label 5. So, totally 5 characters were misclassified in ABPN with Shannon activation function usingNguyen-Widrow.

Table 1 shows performance of the ABPN with Shannon activation function and proposedABPNwith Shannon using Nguyen-Widrow weight generation technique. ABPN with Shannon activation function using Nguyen-Widrow achieved 96% of classification accuracy in 190 epochs whereas ABPN with Shannon activation function achieved 94% of classification accuracy in 410 epochs with the minimized MSE 0.0003. From the Table 1, it has been proved that the ABPN with Shannon activation function using Nguyen-Widrow weight Initialization technique works better for Tamil Palm leaf manuscript character classification than the ABPN with Shannon activation function.

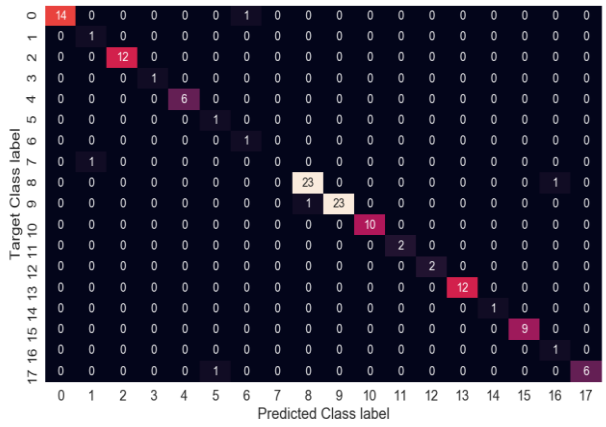

Figure 7. Confusion matrix for ABPN with Shannon activation function with Nguyen-Widrow weight initialization.

V. CONCLUSION AND FUTURE SCOPE

In the proposed work, Tamil palm leaf manuscript characters were segmented using contour based convex hull bounding box segmentation. Then from the segmented characters, 18 labelled dataset were selected which contain totally 130 characters. Those character images were binarized using AdaptiveMean threshold method and the intensity value of pixels of characterwere fed into neural network. The neural networks are used for comparison are: ABPN with Shannon activation function using Nguyen-Widrow weight generation and ABPN with Shannon activation function. The confusion matrix was drawn to visualize the character classification to identify which network worked well for Tamil Palm leaf manuscript character classification. From these two networks, ABPN with Shannon activation function using Nguyen-Widrow weight initialization network achieved 96% of accuracy and gives promising result whereas ABPN with Shannon activation function achieved 94% of accuracy.

REFERENCES

[1] Amit Choudhary, Rahul Rishi, Savita Ahlawat, "Off-Line Handwritten Character Recognition using Features Extracted from Binarization Techniques", *AARSI Procedia, ,* Vol.4, pp.306-312, 2013.

[2] Dino Neinhold, Kilian Schwab, Rolf Dornberger and Thomas Hanne, "Effects of Weight Initialization in a Feedforward Neural Network for Classification Using a Modified Genetic Algorithm", IEEE, pp.6-12, 2015.

[3] Hossin. M., Sulaiman. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations", *International Journal Of Data Mining & Knowledge Management Process,* Vol.5, No.2, pp.1-11, 2015.

[4] JeyaseeliSubavathi S, Kathirvalavakumar T, "Adaptive modified backpropagation algorithm based on differential errors", *International Journal of Computer Science, Engineering and Applications (IJCSEA),*; Vol.1, pp. 21-34, 2011.

[5] Karthigaiselvi. M, Kathirvalavakumar. T, "Structural Run Based Feature Vector to Classify Printed Tamil Characters using Neural Network", *International Journal of Engineering Research and Application,* Vol.7, Issue.7, pp.44-63, 2017.

[6] KavithaSubramani, Murugavalli. S, "A Novel Binarization Method for Degraded Tamil Palm Leaf Images", *IEEE,* pp.176-181, 2016.

[7] Kiruba. B, Nivethitha. A, Vimaladevi. M, "Segmentation of Handwritten Tamil Character from Palm Script using Histogram Approach", *International Journal of Informative and Futuristic Research,* Vol.4,Issue.5, pp.6418-6424,2017.

[8] NarahariSastryPanyam, Vijaya Lakshmi. T.R, Ramakrishnan Krishnan, Koteswara Rao. N.V, "Modeling of Palm Leaf Character Recognition System using Transform based Techniques", *Pattern Recognition Letters,* Vol.84, pp.29-34,2016.

[9] Nibaran Das, Jagan Mohan Reddy, Ram Sarkar, SubhadipBasu, MahantapasKundu, MitaNasipuri, Dipak Kumar Basu, "A Statistical-Topological Feature Combination for Recognition of Handwritten Numerals", *Applied Soft Computing* ,Vol.12, pp.2486-2495 , 2012.

[10] Sornam.M, Muthu SubashKavitha, Poornima Devi.M, "An Efficient Morlet Function based Adaptive Method for Faster Backpropagation for Handwritten Character Recognition", *IEEE,* pp.135-139, 2016.

[11] Sornam.M, Poornima Devi. M, "Tamil Palm Leaf Manuscript Character Segmentation using GLCM Feature Extraction", *International Journal of Computer Science and Engineering,* Vol.6, Issue.4,pp.167-173,2018.

[12] Sornam. M, C. Vishnu Priya, "Deep Convolutional Neural Network for Handwritten Tamil Character Recognition Using Principal Component Analysis", *Smart and Innovative Trends in Next Generation Computing Technologies,* pp. 778-787, 2018.

[13] Tien-Chien Chang and Shu-Yuan Chen, "Character Segmentation using Convex hull Techniques", *International Journal of Pattern Recognition and Artificial Intelligence,* Vol.13, No.6, pp.833-858,1999.

[14] Vellingiriraj. E.K, Balasubramanie. P, "Recognition of Ancient Tamil Handwritten Characters in Historical Documents by Boolean Matrix and BFS Graph", *International Journal of Computer Science and Technology,* Vol.5, pp.65-68,2014.

[15] Vijaya Lakshmi. T.R, "Reduction of Features to Identify Characters from Degraded Historical Manuscripts", *Alexandria Engineering Journal*, pp.1-7, 2017.
[16] YapingZang, Shan Liang, ShuaiNie, Wenju Liu, Shouye Peng, "Robust Offline Handwritten Character Recognition through Exploring Writer-Independent features under the guidance of Printed data", *Pattern Recognition Letters*, Vol.106, pp.20-26,2018.
[17] Yi-Chou Wu, Fei-Yin, Cheng-Lin Liu, "Improving Handwritten Chinese Text Recognition using Neural Network Language Models and Convolutional Neural Network Shape Models", *Pattern Recognition*, Vol.65,pp.251-264, 2017.

Authors Profile

*Dr M. Sornam*received her MSc in Mathematics from the University of Madras in the year 1987, Master's Degree in Computer Applications from the University of Madras in the year 1991 and received her Ph.D from the University of Madras in the year 2013. Since 1991–1996, she worked as a Lecturer in Computer Science at Anna Adarsh College, Chennai. Later, from 1996 to 2000 she worked as a Lecturer in Computer Science at T.S. Narayanasami College of Arts and Science, Chennai. Since 2001, she has been working in the Department of Computer Science, University of Madras. At present, she is working as an Associate Professor in Computer Science at the University of Madras. Her area of interest includes artificial intelligence and artificial neural networks , image processing, data mining, pattern recognition and applications.

Miss. Poornima Devi. M pursed Bachelor of Science in Soka Ikeda College for Women from University of Madras, India in the year 2013, Master of Science in Queen Marys College for Womenfrom University of Madras, India in the year 2015 and Master of Philosophy in University of Madras, India in the year 2016. She is currently pursuing Ph.D. in Department of Computer Science, University of Madras, India since 2017. Her main research work focuses on Artificial Intelligence, Artificial Neural Network, Image Processing, Deep Learning Neural Network.

