**Development of Optimized Ensemble Machine Learning-based Character Segmentation Framework for Ancient Tamil Palm Leaf Manuscripts**

**Introduction**

The recognition of hand written characters were a computer aided system that could read, process and recognition of characters using the specific features of the characters [9]. Character recognition was the method of optical recognition. The recognition could be done online as well as offline. In online character recognition, the flow of stroke was examined and decision was made based on the strokes [10]. But, the offline character recognition was based on processing the features of the optically scanned documents. Palm leaf manuscript was an anciently written document which could be recognized with offline character recognition [11]. The offline recognition involved the enhancement of the input image, character separation, feature extraction and feature recognition. Tamil language has 247 letters that contains 12 vowels, 18 consonants, 216 complex letters and one special character called’ aaytham’. Among those groups, the 12 vowels have a common feature.Character segmentation was a pre-processing step in Optical Character Recognition (OCR) [12]. It was the process of separating the word image into individual character images. It was the fundamental and critical step of OCR. The performance of OCR degraded, due to incorrect character segmentation [13].

Segmentation of characters was a complex task in Indian scripts of cursive handwriting style. Character segmentation became difficult with the variation in writing styles, skew variability etc.Segmentation of document images into text-lines and words wasa important step for the document understanding [14]. However, unlike machine-printed documents, the segmentation of handwritten documents was still considered a challenging problem due to (i) irregular spacing between words and (ii) variations of writing styles depending on the person. For the word segmentation [15], document images were first segmented into text-lines [16]. Then, the word segmentation algorithm (for a single text-line) was applied to individual text-lines. Given a single text-line, the conventional word segmentation algorithms consist of two steps: the first step was to extract candidates for inter-word gaps and the next step was to classify the candidates into intra/inter-word gaps. For the candidate generation, a given text-line wasrepresented with a set of super-pixels and their gaps were considered candidates to be classified. Therewas a binary classification problem that assigned a label, where 0 means that the gap is an intra-word gap and 1 indicates it is an inter-word gap [17]. The global/adaptive thresholding was used in the unsupervised learning techniques such as clustering and Gaussian Mixture Model (GMM) were adopted in the scale space selection approach [18].

Artificial neural network is an example of such a system which adjusts the weights of its links from the training patterns. These weights implicitly work as features for classification. Another example of non-explicit feature-based method is the hidden Markov model [19]. Statistically derived parameters play a vital role in this approach and it needs a very large number of training samples to estimate the probability parameters in a reliable manner [20]. Structural and topological features based tree classifier, and neural network classifiers are mainly used for the recognition of Indian scripts [21]. After the texted document was fed into the system and High Definition (HD) images were obtained, it starts by segmenting every character from the image for identifying the letters. In the next stage, the letters are recognized by the system, and the image was used for the detection of words [22]. Machine Learning Algorithms were applied to go through these processes that are based on the data obtained in the training fees [23]. The machine sends the expected output in a word file format. The system could be easily trained for processing large sets of data that comprise different shapes and styles [24]. Machine learning pleased a crucial role in the segmentation of handwritten documents. It was also beneficial for companies to store necessary documents in a handwritten format [25]. It becomes a lot faster in easier to complete with the help of such technologies.

**Related works**

In 2022, Kavitha and Srimathi [1] have proposed the state of the art CNN in recognizing handwritten Tamil characters in offline mode. CNNs differ from traditional approach of Handwritten Tamil Character Recognition (HTCR) in extracting the features automatically. It has used an isolated handwritten Tamil character dataset developed by HP Labs India. Ithave developed a CNN model from scratch by training the model with the Tamil characters in offline mode and have achieved good recognition results on both the training and testing datasets. This works was an attempt to set a benchmark for offline HTCR using deep learning techniques. These works have produced a training accuracy which was far better compared to the traditional approaches.

In 2021, Lincy and Gayathri [2] have proposed and developed a novel Tamil Handwritten Character Recognition approach by following two major processes, viz. pre-processing and recognition. The pre-processing phase enclosed the RGB to grayscale conversion, binarization with thresholding, image complementation, morphological operations, and linearization. Subsequently, the pre-processed images after linearization were subjected to recognition via an optimally configured CNN. More particularly, the fully connected layer and weights are fine-tuned by a new Self Adaptive Lion Algorithm (SALA) that was the conceptual improvement of the standard Lion Algorithm (LA). The performance of the proposed work was compared and proved over other state-of-the-art models with respect to certain performance measures.

In 2020, Athisayamani and Singh *et al*. [3] have proposed the recognition of B-spline curves to recognize the twelve vowels in the palm leaf manuscripts. The advantage of B-spline curve was uniqueness and robustness. Each vowel in Tamil language has one or more curves of different angle. Recognition of the combination of curves was used to recognize the vowels. The results showed that the proposed method would recognize characters written by different narrators and the proposed method was performed better than other state- of - the art methods in terms of recognition accuracy.

In 2019, Kowsalya and Periasamy [4] have proposed effective Tamil character recognition. The proposed methods have four main process such as preprocessing process, segmentation process, feature extraction process and recognition process. For preprocessing, the input image was fed to Gaussian filter, Binarization process and skew detection technique. Then the segmentation process was carried out, here line and character segmentation were done. From the segmented output, the features were extracted. After that the feature extraction, the Tamil character were recognized by means of optimal artificial neural network. Here the traditional neural network was modified by means of optimization algorithm. In neural network, the weights were optimized by means of Elephant Herding Optimization. The performance of the proposed method was assessed with the help of the metrics namely Sensitivity, Specificity and Accuracy. The proposed approach was experimented and its results are analyzed to visualize the performance.

In 2019, Gupta and Bag [5] have proposed a approach based on the polygonal approximation of the word, which worked on more than one Indian languages. This works depicted the novel approach for script independent character segmentation of handwritten text utilizing basic structural properties of the languages. Digitally Straight Line Segments (DSS) of the word was obtained by applying Polygonal approximation to the word. The segmentation of character was language independent and worked considerably with skew words as well. Experiments were carried out with four popular Indian languages, Hindi, Marathi, Punjabi, and Bangla. The average success rate for character segmentation of four languages was which was satisfactory compared with other existing methods. This used shadow and cumulative stretch feature set with random forest, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and CNN classifiers for character recognition. On experimentation, it was observed that this proposed method provided good accuracy for character segmentation and recognition.

In 2015, Ryu and Koo [7] have formulated the word segmentation problem as a binary quadratic assignment problem that considered pair wise correlations between the gaps as well as the likelihoods of individual gaps. Structured SVM framework proposed method was worked well regardless of writing styles and written languages without user-defined parameters. Experimental results on handwriting segmentation databases showed that proposed method achieved the state-of-the-art performance on Latin-based and Indian languages.

In 2010, Shanthi and Duraiswamy [6] have proposed a system for recognizing offline handwritten Tamil characters using Support Vector Machine (SVM). Data samples were collected from different writers on A4 sized documents. They were scanned using a flat bed scanner at a resolution of 300 dpi and stored as gray-scale images. Various preprocessing operations were performed on the digitized image to enhance the quality of the image. Pixel densities were calculated for 64 different zones of the image and these values were used as the features of a character. These features were used to train the SVM. The SVM was tested for the first time to recognize handwritten Tamil characters. The system has achieved very good recognition accuracy on the handwritten Tamil character database.

In 2008, Surinta and Chamchong [8] have proposed an image segmentation of historical handwriting from palm leaf manuscripts. The process was composed of three steps: 1) background elimination to separate text and background by Otsu‘s algorithm 2) line segmentation and 3) character segmentation by histogram of image. The end result was the character’s image. The results from this research may be applied to optical character recognition (OCR) in the future.

**Problem Definition**

Character segmentation forTamil palm leaf manuscript is highly required due to certain factors like difficult to write and recognize, errors during writing, noise, inaccuracy and overfitting. Consequently, efficient algorithms are developed in the existing methods that are discussed in Table 1. In Structured SVM [6] cost function are more compact and it can be easily extended to other database. But, it does not work, if it has same properties in inter/intra word gaps and super pixel is used to reduce the contents, languages and scanning activities. CNN [2] is used to perform optimum threshold and helps in handling the image handwritten documents clearly. But, it helps in designing the gray image and enhances the visibility and it is complex to recognize the handwritten and the quality of the image is poor, because it seeps ink from the other side of paper also it does not provide accurate recognition for Tamil characters. CNN [5] it is able to process independent handwritten language and provides good precision for the character segmentation. But, it does not encode the position and orientation of image and also in binary process; if stroke is lost then the shape of the character will be inaccurate. B-spline curve [3] it is very unique and it provides robustness also supports in rotating, scaling and extending transformations. But, it is hard to achieve the accuracy in handwritten recognition and it needs more number of curves for specific recognition. CNN [1] it avoids over training of the network and provides good recognition in both testing and training accuracy. But, it gives error during writing, because of same characters also it leads to over fitting problems. Otsu’s Algorithm [8] is used to perform automatic image thresholding and reduces the noise from the scanning process. But, while we convert the image into gray, it depends on the sensitivity and due to workload, it has some limitations when it comes to handwritten. Neural network [4] it has reduced errors while writing and it attains optimal weight and produces high performance and effective recognition rate. But, character editing is time consuming and provides lower accurateness also it is difficult to recognize the handwritten in Tamil character. SVM [7] the size of zone is small, so that it captures more pixel variation and it achieves good recognition accuracy. But, it does not provide robustness and errors are occurred during writing, due to same shaped characters. Hence these challenges inspired us to develop character segmentation for ancient Tamil palm leaf manuscript using machine learning techniques.

**Table 1:** Features and challenges of the existing Tamil character segmentation methods

|  |  |  |  |
| --- | --- | --- | --- |
| **Author [citation]** | **Methodology** | **Features** | **Challenges** |
|  |  |  |  |
| Kavitha and Srimathi [1] | CNN | * It avoids over training of the network. * It provides good recognition in both testing and training accuracy. | * It gives error during writing, because of same characters. * It provides overfitting problems. |
| Lincy and Gayathri [2] | CNN | * It performs optimum threshold and helps in handling the image handwritten documents clearly. * It helps in designing the gray image and enhances the visibility. | * It is complex to recognise the handwritten and the quality of the image is poor, because it seeps ink from the other side of paper. * It does not provide accurate recognition for Tamil characters. |
| Athisayamani, Singh *et al*. [3] | B-spline curve | * It is very unique and it provides robustness. * It supports in rotating, scaling and extending transformations. | * It needs more number of curves for specific recognition. * It is hard to achieve the accuracy in handwritten recognition. |
| Kowsalya and Periasamy [4] | Neural network | * Errors have reduced while writing and it attains optimal weight. * It produces high performance and effective recognition rate. | * Character editing is time consuming and provides lower accurateness. * It is difficult to recognise the handwritten in Tamil character. |
| Gupta and Bag [5] | CNN | * It can able to process independent handwritten language. * It provides good precision for the character segmentation. | * It does not encode the position and orientation of image. * In binary process, if stroke is lost then the shape of the character will be inaccurate. |
| Ryu and Koo [6] | Structured SVM | * The cost function is more compact. * It can be easily extended to other database. | * It does not work, if it has same properties in inter/intra word gaps. * Super pixel reduces the contents, languages and scanning activities. |
| Shanthi, and Duraiswamy [7] | SVM | * When the size of zone is small, it captures more pixel variation. * It achieves good recognition accuracy. | * It does not provide robustness. * Errors are occurred during writing, due to same shaped characters. |
| Surinta and Chamchong [8] | Otsu’s Algorithm | * It is used to perform automatic image thresholding. * It reduces the noise from the scanning process. | * When we convert the image into gray, it depends on the sensitivity. * Due to workload, it has some limitations when it comes to handwritten. |

**Research Methodology**

Tamil is one of the classic languages in the world. The ancient Tamil documents on medicines, literatures, drama, astrology and many other ancient contents were written in stones and palm leaf. The character recognition of palm leaf manuscript is one of the open problems to be solved. In recent data, Optical Character Recognition (OCR) systems have laid hands in the field of most popular language recognitions. Unlike other languages, the Tamil language is more complex to recognize, and hence considerable efforts have been laid in literature. However, the models are not yet well-organized for precise recognition of Tamil characters. Character segmentation plays an important role in developing OCR for handwritten languages. The exactness of character segmentation is the integral factor of OCR. Therefore, an efficient character segmentation approach for historical handwriting from palm leaf manuscripts is mainly developed in this paper for making the ancient Tamil character recognition to be more accurate. The developed model will be enclosed with four steps: 1) Image Collection, 2) background elimination with Image pre-processing, 3) line segmentation and 4) character segmentation. Initially, the images of Tamil palm leaf manuscripts will be collected from the standard online datasets. The collected images of Tamil manuscripts will be pre-processed with the Optimal Binary Thresholding, filtering and Morphological Operation. The pre-processed images will be used for Line segmentation by utilizing the Projection Profile method. The line segmented images will be given to the feature extraction process, where the features from the images like Local Tetra Pattern (TLP), Local Binary Pattern (LBP), Autoencoder features and the deep features will be extracted. These extracted features will be subjected to the character segmentation with the utilization of ensemble machine learning structures. Here, the ensemble machine learning model will be constructed with Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Bayesian Learning (BL) classifiers in order to provide promising results over character segmentation. Moreover, the optimization will be done for making process to be less complex by tuning the parameters in segmentation techniques using the hybrid optimization algorithm of Sandpiper Optimization Algorithm (SOA) [26] and Fireworks Algorithm (FA) [27]. The experimental analysis will be made to declare the efficiency of the developed Tamil character segmentation approach by comparing with other conventional algorithms. The architecture of the developed Tamil character segmentation framework is given in Figure 1.

Image collection

Image pre-processing

DevelopedSOA+FA

Line segmentation

Feature extraction

Character segmentation

Ensemble learning

ANN

SVM

BL

Segmented results

**Figure 1**: Architecture of developed framework

**Expected Outcome**

The proposed model will be evaluated in Python and the performance analysis will be carried out. Here, Type I measures are positive measures like Accuracy, Sensitivity, Specificity, Precision, Negative Predictive Value (NPV), F1Score and Mathews correlation coefficient (MCC), and Type II measures are negative measures like False positive rate (FPR), False negative rate (FNR), and False Discovery Rate (FDR).

**References**

1. Kavitha B.R and Srimathi C., "Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks," Computer and Information Sciences, vol. 34, Issue 4, pp. 1183-1190, April 2022.
2. Lincy, R.B and Gayathri, R. "Optimally configured convolution neural network for Tamil Handwritten Character Recognition by improved lion optimization model," Multimedia Tools and Applications, vol. 80, pp.5917–5943, 2021.
3. Suganya Athisayamani, A. Robert Singh Dr. and T.Athithan Dr." Recognition of Ancient Tamil Palm Leaf Vowel Characters in Historical Documents using B-spline Curve Recognition," Procedia Computer Science, vol. 171, pp. 2302-2309, 2020.
4. Kowsalya, S and Periasamy, P.S., "Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization," Multimedia Tools and Applications, vol.78, pp.25043–25061, 2019.
5. Gupta, D and Bag, S., "Handwritten multilingual word segmentation using polygonal approximation of digital curves for Indian languages," Multimedia Tools and Applications, vol. 78, pp.19361–19386, 2019.
6. Jewoong Ryu and Hyung Il Koo, "Word Segmentation Method for Handwritten Documents based on Structured Learning," IEEE Signal Processing Letters, vol. 22, NO. 8, Aug 2015.
7. Shanthi, N and Duraiswamy, K.," A novel SVM-based handwritten Tamil character recognition system" Pattern Analysis and Applications, vol.13, pp.173–180, 2010.
8. Olarik Surinta and Chamchong "Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts," Intelligent Information of Computer Science, 19 October 2008.
9. Sundaram, S. and Ramakrishnan, A.G. "Performance enhancement of online handwritten Tamil symbol recognition with reevaluation techniques," Pattern Analysis and Applications, vol. 17, pp.587–609, 2014.
10. Raghavan, S.V., Vasanth Kumar Mehta, R. and Srinivasu, G. "Palm leaf manuscripts: Saaswathaiswaryam the eternal treasure of India," CSIT 9, pp.31–36,2021.
11. M. Mohamed Sathik and R. SpurgenRatheash,"Optimal Character Segmentation for Touching Characters in Tamil Language Palm Leaf Manuscripts using Horver Method," Blue Eyes Intelligence Engineering & Sciences Publication, pp. 2278-3075, vol. 9 Issue-6, April 2020.
12. Suganya, T.S. andMurugavalli, S.,"A hybrid group search optimization: firefly algorithm-based big data framework for ancient script recognition," Soft Computing, vol. 24, pp.10933–10941, 2020.
13. Gayathri Devi S, SubramaniyaswamyVairavasundaram, Yuvaraja Teekaraman, Ramya Kuppusamy and Arun Radhakrishnan, "A Deep Learning Approach for Recognizing the Cursive Tamil Characters in Palm Leaf Manuscripts", Computational Intelligence and Neuroscience, vol. 2022, pp.15 , 2022.
14. MerlineMagrina M and Dr. Santhi M, "Ensemble Classifier System for OfflineAncient Tamil Character Recognition, SSRG International Journal of Electronics and Communication Engineering (SSRG-IJECE), Mar 2019.
15. R. Jayakanthan, A. Hiran Kumar, N. Sankarram, B. S. Charulatha and Ashwin Ramesh, "Handwritten Tamil Character Recognition Using ResNet," International Journal of Research in Engineering, Science and Management,vol.3, Issue-3, March-2020.
16. A. Nivethitha and Dr. M. Vimaladevi, "Segmentation of Handwritten Tamil Character from Palm Script using Histogram Approach," Computer Science, 2017.
17. Poornima Devi.M and Sornam.M, "Classification of Ancient Handwritten Tamil Characters on Palm Leaf Inscription Using Modified Adaptive Back propagation Neural Network with GLCM Features," ACM Transactions on Asian and Low-Resource Language Information Processing, vol.19, pp 1–24, November 2020.
18. Poornima Devi. M and M. Sornam "Systematic Classification of HistoricalHandwritten Tamil Palm Leaf Manuscript using CART algorithm and RBF Network," International Journal of Innovative Technology and Exploring Engineering (IJITEE),pp. 2278-3075, vol.11, September 2019.
19. AbiramiMurugappan, "Junction Point Elimination based Tamil Handwritten Character Recognition: An Experimental Analysis,” Journal of Systems Science and Systems Engineering, December 2019.
20. Kandasamy Shanmugam, "Newton Algorithm Based DELM for Enhancing Offline Tamil Handwritten Character Recognition," International Journal of Pattern Recognition and Artificial Intelligence, April 2022.
21. Ahlawat Savita, Amit Choudhary, and Anand Nayyar,Saurabh Singh, "Improved Handwritten Digit Recognition Using Convolution Neural Networks (CNN)," Sensors, June 2020.
22. R. S. Sabeenian, M. E. Paramasivam, R. Anand and P. M. Dinesh, "Palm-Leaf Manuscript Character Recognition and Classification Using Convolutional Neural Networks," Computing and Network Sustainability, 2019.
23. AbiramiMurugappan and M. Antony Robert Raj, “Structural representation-based off-line Tamil handwritten character recognition," Soft Computing, vol.13, pp. 24:1447–1472, January 2020.
24. K. S. Hemanth and M. Malini, "Survey on Handwritten Characters Recognition in Deep Learning," Ubiquitous Intelligent Systems, pp.123-133, January 2022.
25. Nasira, G. M., and P. Banumathi. "Off-Line Handwritten Character Recognition with Hidden Markov Models." Artificial Intelligent Systems and Machine Learning, pp.57-61,2011.
26. Amandeep Kaur, Sushma Jain and Shivani Goel,"Sandpiper optimization algorithm: a novel approach for solving real-life engineering problems,"Applied Intelligence, vol. 50, pp. 582–619, 2020.
27. Ying Tan and YuanchunZhu,"Fireworks Algorithm for Optimization,"Advances in Swarm Intelligence, First International Conference, June 12-15, 2010.