Academic Research Journal of Science and Technology (ARJST), (ISSN: 3048-9644)
Peer Reviewed Multidisciplinary Monthly Scholarly International Journal
Vol. 1 No. 07, March 2025

Design and Development of Enhanced Deep Learning Methodology for Tamil Manuscripts Extraction using hybrid CNN-LSTM-CTC

Dr P. Jayapriya
Assistant Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College,Pollachi,India. topriyadamu@gmail.com
8973006779

Abstract

Extraction of data from manuscripts using Deep Learning emerged as an intriguing task in fields like historical document analysis, records digitization and information retrieval. With the emergence of deep learning, new potential methods for efficient and accurate data extraction have surfaced. This paper discovers a new systematic approach with the combination of segmentation, Convolutional Neural Network (CNN) and classification techniques. This helps to dig out meaningful information from handwritten and printed manuscripts. Moreover tamil manuscripts, rich in cultural and historical significance faces distinctive challenges. This includes factors like complex nature, varied writing styles and deprivation over time. So it is necessary to review the challenges, datasets, pre-processing methods to perform this work deep learning approaches. A hybrid CNN-Long Short-term Memory (LSTM)- Connectionist Temporal Classification(CTC) is proposed to empower all challenges. In this approach, firstly preprocessing can be done using Optical Character Recognition (OCR).Secondly a well-suited LSTN-CNN is used for capturing sequential dependencies in text. Finally CTC function helps in handwritten text recognition and text extraction of Tamil manuscripts.

Keywords: Deep Learning, Tamil manuscripts, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Connectionist Temporal Classification (CTC)

1. Introduction

Manuscripts cover valuable information across various domains such as history, medicine, and architectural design of sculptors. This valuable information need to transform in to digital form. Sometimes manual transcription and analysis are labor-intensive and error-prone. So a technology with deep learning techniques, particularly in conjunction with machine learning and deep learning, provides an automated solution for processing such documents. This research focuses the development of new architecture using deep learning. It acts as a pipeline to extract text and structured data from manuscripts. The primary goal of above method is to detect Tamil cursive letters imprinted in palm leaves. A proposed character recognition method is used along with the flexible B-spline curve to identify different tamil vowel letters [1]. This is accomplished by categorizing the characters into relevant category based on characteristics like shape and size. The steps like i) Image Preprocessing ii) Feature Extraction (iii)Segmentation (iv) Classification using CNN will assist to achieve greater accuracy.

2. Related works

Islam et al. [2] proposed a Convolutional Neural Network (CNN) model for manuscript characters classification by classifying the Beowulf manuscript's Old English characters. This approach may help machine learning researchers in automated character recognition of ancient manuscripts, which open a new horizon for researchers in history, arts, literature, science, and technology. To perform the classification task, developed a CNN model to train and test with the Beowulf manuscript's dataset. Moreover, comparative studies are performed using other Machine Learning (ML) models such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and XGBoost. Also used pre-trained models such as VGG16, MobileNet, and ResNet50 to extract features from the Beowulf manuscript character images dataset.Then used those features to train the SVM, KNN, DT, RF, XGBoost models, and subsequently tested these models with Beowulf test dataset. And recorded the recognition accuracy results and evaluated each model performance by computing recall, precision, and F1 scores values.

Sinthuja et al. [3] proposed an approach to extract the text from image including hand-written and normal text by using Convolutional Neural Networks (CNN) and Bi-directional Long Short-term Memory (BiLSTM). This hybrid architecture of combining CNN and BiLSTM, focus on local patterns. It performs good at capturing context and long-range dependencies.BiLSTMs are well suited for tasks that require understanding of the text's sequential nature. The model achieved accuracy of 88.58 for hand written text and also achieved accuracy of 90.8 for normal text.

Daniel et al. [4] proposed OCR model with screen readers. This system performs well when images contain texts printed on plain backgrounds. So OCR system that can perform well on images with texts on complex background becomes a fundamental necessity. This work has attempted to address the issue by designing an algorithm that leverages some pitfalls. Already existing OCR system helps to improve its performance in images with complex backgrounds.

Geetha et al. [5] proposed a model that makes use of Convolutional Neural Network (CNN) with Long Short Term Memory (LSTM) method to detect and identify the text. The proposed model has been trained and tested using MNIST dataset consisting of full page images from

3600 writers and 800,000 images. Experimented with different number of epochs, it is found that the proposed model is able to identify the text with an accuracy of about 88.8%.

Liang et al. [6] proposed a DC CNNLSTM model, which incorporates multiplication by weights (w). Experiments results have shown that DC CNN-LSTM can effectively distinguish the emotional level of different words in sentences, and assign different learning weights to different words, so that it can learn the sentiment features of each word in a differentiated way.

Jailingeswari et al. [7] proposed a ThimuNet CNN model to classify cursive Tamil characters, and its performance is confirmed. In order to improve the performance of proposed model, the weight of the input parameters must be optimized, which is considered as future scope of this research model. In addition, preprocessing plays a major role in removing unwanted cursive writings; therefore, an effective preprocessing technique is implemented along with the developed model for better performance.

Sivan et al. [8] proposed an intelligent character segmentation model coupled with deep learning-based recognition. The main focus of this work is to create an automated Manuscript reader that can work efficiently for real-time Archaeological and historical applications. The Augmented HPP method proposed helps to handle segmentation cases that were previously not considered by existing segmentation techniques. The novel Punch Hole removal algorithm effectively locates and removes the Punch hole impressions in the manuscript images. An automated content cropping algorithm is also introduced to reduce the manual work required in real-time applications. Finally, Tamil Character Recognition is performed by Modified CNN with 125 classes. A significant feature of this CNN is that it can recognize all 247 letters and 12 numeric characters in Tamil language with the limited 125 classes. It significantly reduces the complexity of the Network. While analyzing the performance of the proposed intelligent Tamil character recognition model, the system attained an average segmentation accuracy of 98.25%, a recognition accuracy of 96.04%, and a loss of 0.21%. Methodology

3. Proposed Methodology

In manuscripts, preprocessing helps to improve accuracy with the help of Optical Character Recognition (OCR).This can be done through the ocr() function that helps to extract text from images. OCR enhance the recognition process by adjusting the image quality, using binarization.This process can be applied using morphological operations like dilation (imdilate()) and erosion (imerode()). Training of OCR can be done with Convolutional Neural Networks (CNNs) [9].

The OCR preprocessed manuscript images can be converted into text sequences using Long Short-Term Memory (LSTM).Training LSTM model helps to recognize these sequences, and this model is used for predictions or classifications. Recurrent layer of LSTM helps to recognize sequences of Handwritten Text images.Many reviews provides that valuable insights with various aspects of OCR, that drives to the further advancements [10].The core objective of Long Short Term Memory (BiLSTM) model is to reduce manual paper checking and enhance the accuracy of paper scanning and text recognition using CRNN (Convolutional Recurrent Neural Networks)[11].

Connectionist Temporal Classification (CTC) is a loss function used for sequence-to-sequence tasks where the alignment between input and output sequences is unknown. CTC is useful in recognition of tamil text and the model needs to learn the correct sequence of outputs (characters or words) without precise alignment. A hybrid CNN-LSTM-CTC model architecture typically consists of a convolutional Neural Network (CNN) for preprocessing of collected manuscript images using OCR.Preprocessed images undergoes segmentation and classification using LSTM.Finally recognition and extraction using Connectionist Temporal Classification (CTC) loss function helps to align the predicted labels with the input sequence as shown in Figure 1.

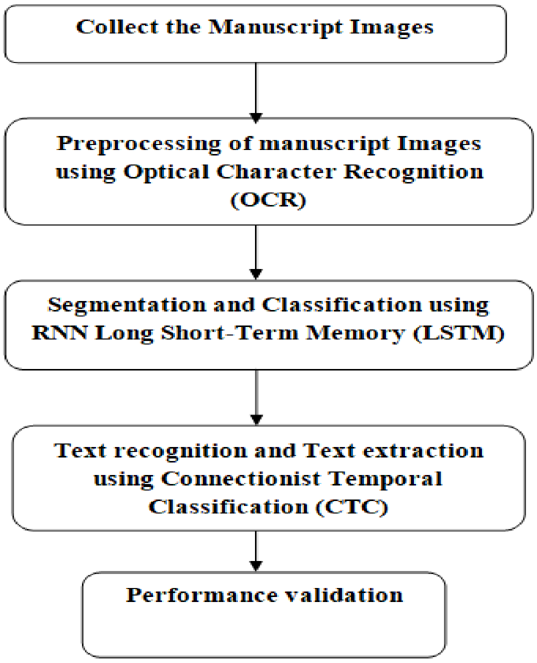
Collect the Manuscript Images
Preprocessing of manuscript Images using Optical Character Recognition (OCR)
Segmentation and Classification using RNN Long Short-Term Memory (LSTM)
Text recognition and Text extraction using Connectionist Temporal Classification (CTC)
Performance validation

Figure 1: Proposed Architecture of hybrid CNN-LSTM-CTC model

3.1. Data Preparation

The images of tamil manuscripts written in stone, inscriptions are captured from temples using camera. Sometimes these images may be dull due to the lighting conditions and may contain lot of noise. So the dataset is created from hp labs (hpl) handwritten tamil characters. The success of a deep learning model depends heavily on the quality and quantity of data.

3.2. Preprocessing the Manuscript Images

Manuscripts were collected from libraries, online archives and synthetic datasets. U-Net-based semantic segmentation model as shown in Figure 2 was trained to identify lines, words, and characters.

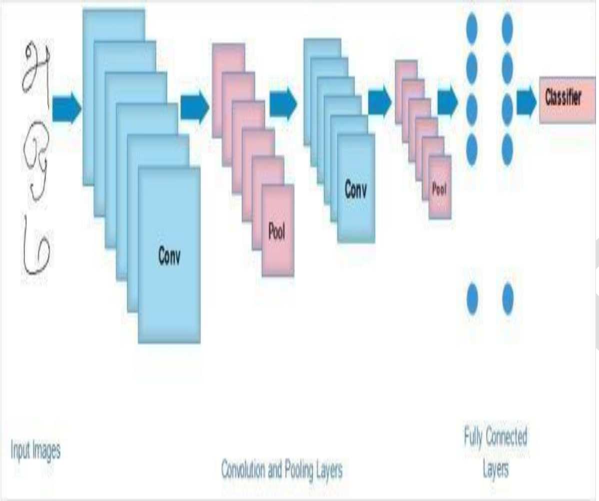

Figure 2: U-Net CNN model

Image Enhancement functions like imadjust(),histeq() and medfilt2() are used to carry out process like Histogram equalization, contrast adjustment and noise reduction to improve the clarity of the manuscript images.

Converting in to binary images helps to isolate the text from background. Functions like imbinarize() or adaptthresh() are used, and is often essential for Optical Character Recognition (OCR).The imrotate() function or Hough transform (hough() and houghpeaks()) helps to correct skew and fix any misalignment in manuscript images.

3.3. Text Recognition with OCR

Optical Character Recognition (OCR) is performed through the ocr() function, which allows to extract text from images. OCR Preprocessing enhance the recognition process by adjusting the image quality, using binarization, and applying morphological operations like dilation (imdilate()) and erosion (imerode()).

Algoritmn

Preprocess:

Step 1: Load the image using imread('manuscript.jpg').

Step 2: Preprocess image using contrast adjustment (imadjust) and noise filtering (medfilt2)

Step 3: Apply OCR function to extract results

Step 4: Apply morphological operations

Post-process

Text: Clean the extracted text for further analysis.

3.4. Training OCR Model:

If the handwriting or font used in the manuscripts is unique, train OCR model using machine learning capabilities. This can be done using deep learning frameworks like Convolutional Neural Networks (CNNs).

Step 1: Input image

Step 2: Perform batch normalization using binarization and noise removal in CNN

Step 3: Use activation function like rectified linear units (relu)

Step 4: Perform maxpooling

Step 5: Text Region Detection and Text Recognition

Step 6: Recognize the text within the detected regions

Step 7: Identify match classes using LSTM

Step 8: Entity Recognition

Step 9: Extract meaningful entities (dates, names, etc.) from the recognized text.

Post-Processing:

Organize the extracted data into structured formats (e.g., CSV, database).

4. Evaluation Metrics

The performance of system using Sensitivity, Specificity and Accuracy of Data in the dataset are divided in to two classes not pedestrian (the negative class) and pedestrian (the positive class). These performance metrics are calculated using the True Positive (TP),True Negative(TN),False Negative (FN), and False Positive (FP). TP is the number of positive cases that are classified as positive.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

FP is the number of negative cases that are classified as positive.TN is the number of negative cases classified as negative and FN is the number of positive cases classified as negative.The sample manuscript is loaded as shown in Figure 3.

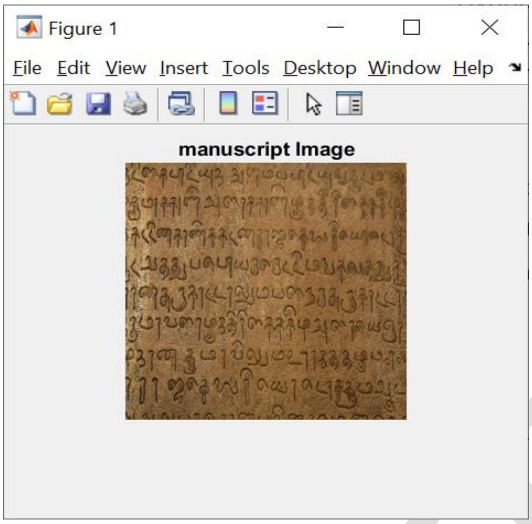

Figure 3: Loading the Input Image

The loaded input image is preprocessed by enhancing contrast as shown in Figure 4.The contrast enhanced image is processed with batch normalization using binary function as shown in Figure 5. This section validates the proposed CNN performance with techniques in terms of various parameter metrics.

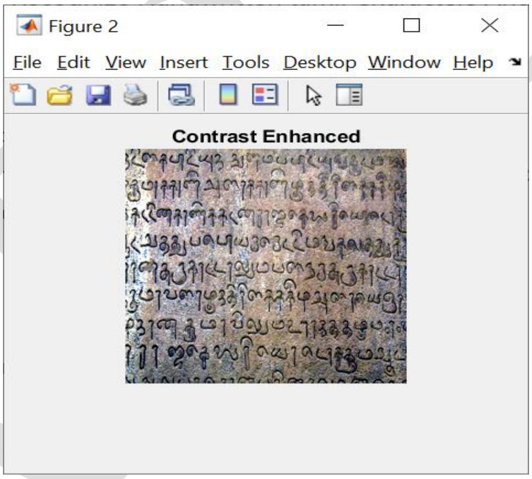

Figure 4: Contrast adjusted Image

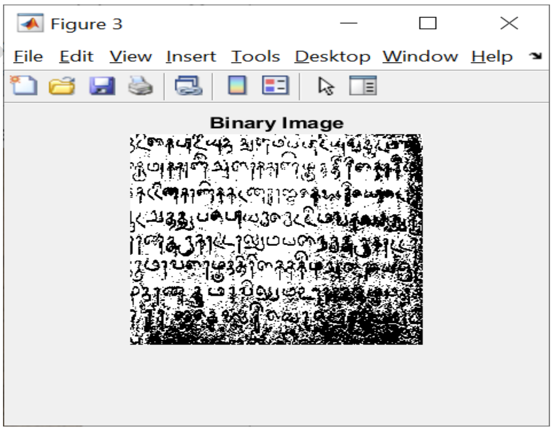

Figure 5: Batch Normalized Image using binarization

The major parameters used are Sensitivity, Specificity and Accuracy. However, in this work, only a few parameters are considered for the validation process; the reason is that it is a collected cursive Tamil Palm Leaf manuscript and cannot apply all major parameters.

Table 1.1 Performance Metrics

Architecture/ Metrics	**LeNet**	**ResNet**	**Hybrid CNN-LSTM-CTC**
Sensitivity (%)	**97.9**	**96.9**	**99.67**
Specificity (%)	**97.8**	**96.84**	**99.88**
Accuracy (%)	**96.87**	**96.45**	**99.78**

The values in table 1.2 clearly prove that the proposed hybrid CNN-LSTM-CTC model achieved higher performance than existing techniques in terms of Sensitivity,Specificity and Accuracy. For example, the LeNet achieved Sensitivity (%) of 97.9, Specificity (%) of 97.8 and 96.87% accuracy, ResNet achieved Sensitivity (%) of 96.9, Specificity (%) of 96.84 with 96.45% accuracy. Hybrid CNN-LSTM-CTC achieved Sensitivity of 99.67 percentages, Specificity of 99.88 percentages and 99.78% accuracy.The Figure 6 represents the comparision of performance metrics of hybrid CNN-LSTM-CTC model.

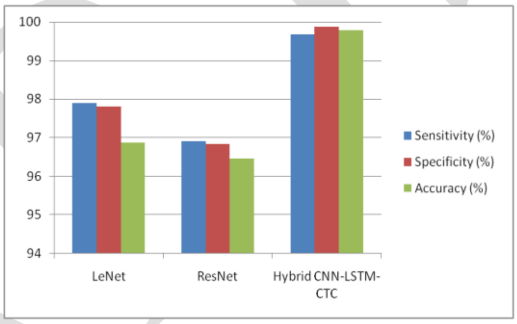

Figure 6: Comparison of Performance Metrics

5. Conclusion and future perspectives

A hybrid CNN-LSTM-CTC model provides an efficient way to recognize Tamil handwritten characters by leveraging CNN for feature extraction,LSTM for sequence modeling and CTC loss for label alignment.This work demonstrates a robust pipeline combining segmentation, CNN-RNN architectures, and classification for efficient data extraction from manuscripts. Future directions include incorporating transformer-based architectures and unsupervised learning to handle unannotated datasets effectively.

References

[1] Suganya Athisayamani, A. Robert Singh, T. Athithan(2020),Recognition of Ancient Tamil Palm Leaf Vowel Characters in Historical Documents using B-spline Curve

Recognition,Procedia Computer Science,Volume 171,Pages 2302-2309,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2020.04.249.

[2] Islam, M. A., & Iacob, I. E. (2023). Manuscripts Character Recognition Using Machine Learning and Deep Learning. *Modelling*, *4*(2), 168-188. https://doi.org/10.3390/modelling4020010

[3] M Sinthuja, Chirag Ganesh Padubidri, Gaddam Sai Jayachandra, Mudduluru Charan Teja, Golthi Sai Pavan Kumar(2024),Extraction of Text from Images Using Deep Learning,Procedia Computer Science,Volume 235,Pages 789-798,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2024.04.075.

[4] Akinbade, D., Ogunde, A. O., Odim, M. O., & Oguntunde, B. O. (2020). An adaptive thresholding algorithm-based optical character recognition system for information extraction in complex images. *Journal of Computer Science*, *16*(6), 784-801.

[5] Geetha, M., Suganthe, R. C., Nivetha, S. K., Hariprasath, S., Gowtham, S., & Deepak, C. S. (2022, January). A hybrid deep learning based character identification model using CNN, LSTM, and CTC to recognize handwritten english characters and numerals. In *2022 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6). IEEE.

[6] Liang, S., Zhu, B., Zhang, Y., Cheng, S., & Jin, J. (2020, December). A double channel CNN-LSTM model for text classification. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (pp. 1316-1321). IEEE.

[7]I.Jailingeswari,S.Gopinathan(2024),Tamil handwritten palm leaf manuscript dataset (THPLMD),Data in Brief,Volume 53,110100,ISSN 2352-3409,https://doi.org/10.1016/j.dib.2024.110100.

[8] R. Sivan, T. Singh and P. B. Pati(2022), "Malayalam Character Recognition from Palm Leaves Using Deep-Learning," *2022 OITS International Conference on Information Technology (OCIT)*, Bhubaneswar, India, pp. 134-139, doi: 10.1109/OCIT56763.2022.00035.

[9]T. M. Saravanan, M. Jegadeesan, P. A. Selvaraj, P. Gopika, R. Kavinesh and G. S. Mahashwetha, "Enhanced Deep Learning Techniques to Classify Tamil Handwritten Characters," *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Krishnankoil, Virudhunagar district, Tamil Nadu, India, 2024, pp. 1-6, doi: 10.1109/INCOS59338.2024.10527587.

[10]Devi S G, Vairavasundaram S, Teekaraman Y, Kuppusamy R, Radhakrishnan A. A Deep Learning Approach for Recognizing the Cursive Tamil Characters in Palm Leaf Manuscripts. Comput Intell Neurosci. 2022 Mar 11;2022:3432330. doi: 10.1155/2022/3432330. Retraction in: Comput Intell Neurosci. 2023 Aug 2;2023:9856274. doi: 10.1155/2023/9856274. PMID: 35310599; PMCID: PMC8933122.

[11]Prabakaran N., Kannadasan R., Krishnamoorthy A., Vijay Kakani(2023),A Bidirectional LSTM approach for written script auto evaluation using keywords-based pattern

matching,Natural Language Processing Journal,Volume 5,100033,ISSN 2949-7191,https://doi.org/10.1016/j.nlp.2023.100033.(https://www.sciencedirect.com/science/article/pii/S2949719123000304)

[12] Alhamad, H. A., Shehab, M., Shambour, M. K. Y., Abu-Hashem, M. A., Abuthawabeh, A., Al-Aqrabi, H., Daoud, M. S., & Shannaq, F. B. (2024). Handwritten Recognition Techniques: A Comprehensive Review. *Symmetry*, *16*(6), 681. https://doi.org/10.3390/sym16060681.

Copyright (c) 2025 Author