Multilingual Character Detection

ARPIT YADAV
Department of CSE
Vardhaman College of Engineering
Hyderabad, India
yarpit2003@gmail.com

GANGA ARCHITH RAJ
Department of CSE
Vardhaman College of Engineering
Hyderabad, India
archithrajganga@gmail.com

SABEEHA FIRDOUS
Department of CSE
Vardhaman College of Engineering
Hyderabad, India
sabeehafirdous03@gmail.com

Dr Y VIJAYALATA
Department of CSE
Vardhaman College of Engineering
Hyderabad, India
vijaya@ieee.org

K NIVEDITHA
Department of CSE
Vardhaman College of Engineering
Hyderabad, India
nivi6k@gmail.com

***Abstract*—The goal of this project is to use deep learning to create a system that can identify and classify handwritten Telugu characters and texts. To improve the model's accuracy in recognizing various Telugu letters, it is trained using a labeled dataset. Instant recognition results will be available for users who upload handwritten characters. The project is designed to serve as both a learning tool and a digitizer of handwritten Telugu scripts. The accuracy of the system can be improved step by step by using user-generated datasets.After completing the research work we have proposed to use a Convolutional Neural Network (CNN) to achieve efficient Telugu character recognition. The gap between handwritten and digital processing can be connected by the project, and its applications can extend to Optical Character Recognition (OCR) for Telugu documents that are scanned.**

***Index Terms*——OCR (Optical Character Recognition), Language Detection, Multilingual OCR, Text Segmentation,Convolutional Neural Networks (CNNs).**

I. Introduction

Multilingual Character Recognition makes up a vital aspect of on-line character recognition and converting handwritten texts into a digital form, so that storage, searching, and processing of information is possible. It is used for various purposes, for example, from reading of old manuscripts to automatic document processing and access to native languages. Though we know that the the task of recognizing complex scripts like English has been done to a greater extent, script recognition is difficult in case of Telugu. The major challenge here is the huge collection of characters, unique ligatures, and different types of handwriting. Telugu is a Dravidian script, which has millions of users and it has curved characters and that is the reason why it has diacritical marks that increase the complexity of the recognition task. Again, in the case of English, as the letters are usually individual, there is no problem with their recognition. Handwriting Telugu, on the other hand, can have variation in stroke width, connection of letters, and presence of characters over each other, making it difficult to segment and recognize correctly. Classic Optical Character Recognition (OCR) systems are usually not accurate by handwritten Telugu text as they are not reliable. That is to say, they are based on the thinking that all users will have different handwriting. The study will propose an advanced deep learning structure for Telugu handwritten text recognition which is designed to achieve these objectives and can be a feasible and optimal solution. The application of Convolutional Neural Networks (CNNs) will allow the system to decide each character and the sequence-based models ,the short and long phrases and finally the sentences thus, the style of the writing will be very similar.In order to improve the model's learning capacity of differnt styles of handwriting, the process will undergo various procedures like character segmentation, noise elimination, and data augmentation.In order for the model to analyze the telugu script or character, the process is done through a lot of techniques such as pre-processing of the images,data training,data testing,using various algorithm for image detection and finally giving the output. This study aims to introduce a very accurate process that will make it possible to read and digitize hand-written Teleugu texts. At the end of this project, a reliable and efficient system will be created, which is a result of the development of various deep learning architectures and a very careful with great attention to every detail evaluation of model performance on the precision metrics along with the confusion matrix analysis. This research will result in the improvement of the efficiency and the increased accessibility of the Telugu handwriting recognition technology and also, it can be used to convert the old manuscripts and various hand-written documents to digitalized format.

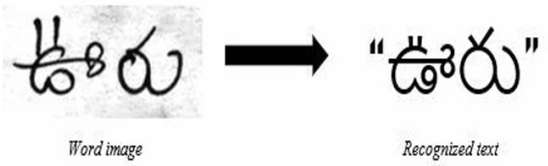

Fig. 1. Handwritten word converting to Digital text

II. Literature Review

Multilingual Character Recognition (HCR) is a crucial area in pattern recognition and optical character recognition (OCR). Telugu, a Dravidian language, has many unique challenges in recognition of the handwritten text due to its complex script structure, connectivity between characters, and high variations in handwriting styles. Over the years, researchers have developed various segmentation techniques, feature extraction methods, and deep learning-based models to improve the performance of the OCR. The earliest efforts in Telugu text recognition were focused on rule-based systems and handcrafted feature extraction methods. Deepa, Ashlin, Guru Sai Jayanth Kalluri, Zeeshan Mohammed, Mantri Pramod Sai Sushank, and Atul Negi implemented zoning and structural feature vectors to recognize isolated Telugu characters and achieving moderate accuracy. [3]

Prakash, K.C.,Srikar, Y.M.,Trishal, G., Mandal, S, Channappayya,S.S implemented a model approach that follows the standard OCR pipeline which is done in step by step method such as skew correction, word segmentation, character segmentation, recognition. [10]

Traditional OCR techniques relied on handcrafted features such as zoning, projection profiles, and shape-based descriptors. Jomy, John, K. V. Pramod, and Balakrishnan Kannan proposed a methodology which improved recognition performance, effective segmentation, and optimized classifier results. [6]

Kumar, P. Pavan, Chakravarthy Bhagvati, Atul Negi, Arun Agarwal, and Bulusu Lakshmana Deekshatulu have developed a model in which after including the proposed algorithms into our OCR system, the previous and the improved OCR systems are compared on around 1000 images . The error rate for the provided page is calculated by using a traditional string matching algorithm, Levenshtein edit distance. [7]

SNS Rajasekaran, BL Deekshatulu have developed A two-stage recognition model for the recognition of the Telugu alphabet (with more than 2000 different characters) is presented. [11]

Meanwhile, SB Madhu, CV Aravinda, MS Sannidhan used two primary approaches used to recognise characters from image composed of Kannada characters that is handwritten by utilising Convolutional Neural Networks and Transfer Learning. [8]

With the development of deep learning, researchers have moved towards CNNs, LSTMs, and hybrid models for handwritten character recognition. Muni Sekhar V Velpuru introduced a deep learning-based OCR system for handwritten Telugu recognition, which outperformed traditional models. Their research used transfer learning and data augmentation techniques to optimize recognition accuracy.

Further, T. Ganji, M. S. Velpuru, and R. Dugyala, implemented the dataset and used a model called VGG-16. By using the VGG-16 they divided the data into testing and training set. They trained and tested the data using VGGNET-16 and achieved good accuracy. [5]

NS Rani, T Vasudev have done the analysis and have taken the single character blocks as simple characters and compound characters, multicharacter blocks are the blocks with more than one character. In every cases related two single and two-three character blocks the analysis has achieved almost 96.56 accuracy. [13]

Buddaraju Revathi, B.N.V., Marapatla, A.D.K., Veeramanikanta, K., Dinesh, K. and Supraja, have done a comprehensive analysis conducted on SqueezeNet, VGG19, and ShuffleNet within the range of image classification which has given distinct performance characteristics. [1]

VL Sravani, PP Singh have used a customized CNN architecture that effectively learned various and distinctive features while combining SVM classifiers with various kernels ,that enhanced classification performance, achieving an accuracy of 99.86 with the RBF kernel. [15]

BVS Rao, JNRBV Venkata, NTKS Rao to improve the training accuracy and to minimize the number of epochs in training phase, a Unicode based HCR (U-HCR) is made. This is used for connecting a scanned handwritten character from Telugu language to English language. [14]

B Meena, KV RAO, S CHITTINENI have done a work that has shown the usage of a Genetic Algorithm (GA) to find the optimal parameter performance for a Convolutional Neural Network. [9]

MS Das, CRK Reddy, K Rahul, A Govardhan have developed a system in which text in multi lingual documents can be recognized .Whereas in conventional OCR systems, text can be recognized only in a particular language . [2]

CS Ram, UTS Abhiroopika, S Chinmayee have used a small dataset of 516 samples and the CNN architecture demonstrated in the paper achieves an accuracy of 79.61. [12]

Dhanikonda, Srinivasa Rao, Ponnuru Sowjanya, M. Laxmidevi Ramanaiah, Rahul Joshi, B. H. Krishna Mohan, Dharmesh Dhabliya, and N. Kannaiya Raja have made a model which when compared to existing algorithms, the suggested method's results are more accurate. Even on a larger dataset, the approach can enhance recognition accuracy . Thee suggested CNN architecture is trained on digital characters. [4]

III. Methodology

The aforementioned Multilingual Character Recognition System is a structured approach and includes data preprocessing, segmentation, feature extraction, deep learning-based classification, and evaluation. Below are the steps of the data preprocessing, the main feature of the article of the reccognition system is to get the optimal outcomes.In total, this section gives us the the methodology used in developing the recognition system that would be able to tell text apart from other objects.

A. Data Preprocessing

The dataset which is used in this research is a written Telugu characters and words by the human hands. To ensure uniformity and improve the model performance, the following preprocessing steps are applied:

a) Image Resizing: To maintain consistency, the input images have all been resized (if they were not at the point of insertion) to the size of 64x64 pixels.

b) Binarization: The otsu's thresholding technique is used to convert images to binary format ,by which we can get perfect contrast needed for the recognition.

c) Morphological Processing: Usage of erosion, dilation, and closing operations allows getting rid of noise and helps in adding to the clarity of strokes for better recognition.

d) Data Augmentation: The data set is exposed to random rotations, scaling, contrast adjustments, and addition of Gaussian noise to fix the problem of generalization.

B. Segmentation Techniques

The main part or principle of Segmentation is the extraction of individual letters. from the handwritten words. The following methods are implemented:

a) Connected Component Analysis (CCA): It identifies different character regions based on the connection between their pixels.

b) Projection Profile Analysis: In this technique the intensity changes in pixels are analyzed, and part of text can be separated into lines and words.

c) Contour-Based Segmentation: This technique applies the OpenCV's bounding box detection to isolate individual characters for recognition purpose.

C. Deep Learning Model Architecture

This new system is being used for the more efficient use of a 'hybrid CNN-RNN architecture', a combination of 'Convolutional Neural Networks (CNNs) for feature extraction' and 'Recurrent Neural Networks (RNNs) for sequence modeling' which means getting to know the sequence of characters.

a) Convolutional Neural Network (CNN) for Feature Extraction : It consists of three convolutional layers. In Telugu, the three convolutional layers recognize spatial patterns. Batch Normalization and Max-Pooling layers feedback work together to improve recognition Dropout layers with a rate of 30

b) Recurrent Neural Network (RNN) for Sequence Learning n: Bidirectional LSTMs (Long Short-Term Memory networks) are the best ways to get rid of the dependencies of handwritten words. The dropout method is a regularization technique which involves randomly ignoring or "dropping out" some layer outputs during training, used in deep neural networks to prevent overfitting which helps in the optimal performance of the system.

c) Connectionist Temporal Classification (CTC) Loss : The CTC loss function is used in this paper so that the network learns logical sequence with a function that does not need character segmentation. Moreover, it enables the model to handle sequences of extremely different lengths,which will help us to recognize various texts of any length.

D. Algorithm for Multilingual Character Recognition

The implemented model uses the real time detection system which is used to deliver the instant feedback.

Below is the Algorithm, Algorithm: Telugu Handwritten Text Recognition

Input : Preprocessed images of characters. Output : Digital Telugu text 1.Preprocess dataset images that include -resizing, noise removal, augmentation. 2.Segment the characters by the usage of contour-based detection and the extraction of bounding boxes. 3.Pass just one of the images made of CNN layers through the system for feature extraction. 4.Process the extracted features using Bidirectional LSTMs for sequence modeling. 5.Using CTC loss function for sequence alignment and text prediction. 6.Optimizing the model using the Adam optimizer and a dynamic learning rate scheduler. 7.Train the model over multiple epochs while validating on the test data. 8.Evaluating the model using accuracy, confusion matrix, and word error rate (WER). 9.Deploy the trained model for real-time recognition of handwritten text.

E. Model Training and Optimization

The model is trained using a large dataset of Telugu handwritten characters and words,applying the following strategies:

a) Optimizer : Adam optimizer with an initial learning rate of 0.001.

b) Learning Rate Scheduling: : The ReduceLROnPlateau technique from TensorFlow arranges the learning rate dynamically and determines rate adjustment based on validation performance.

c) Batch Size and Epochs: Training is conducted with a batch size of 32 and a minimum of 50 to a maximum of 100 epochs, to make sure the convergence.

F. Model Evaluation Metrics

In order to convinve that the recognition system is working well, the following evaluation metrics are used:

a) Test Accuracy : The trained model achieved a test accuracy of 82.74, which is very much effecient when it comes to character recognition.

b) Confusion Matrix Analysis: The confusion matrix is employed to clasify those characters which are missclasified most oftenly. o Top 5 misclassified characters include: "dha" $\rightarrow$ "tha" (8 occurrences) "tha" $\rightarrow$ "dha" (8 occurrences) "am" $\rightarrow$ "aha" (5 occurrences)

c) Word Error Rate (WER) and Character Error Rate (CER) : WER measures text-level error, whereas CER measures character error. The model shows a relatively low CER, which is quite impressive for a single-character recognition.

IV. Implementation and Results

A. Implementation

The Multilingual Character Recognition System was implemented through a structured workflow, involving environment setup, data processing, deep learning model training, evaluation, and real-time testing.The following sections outline the step-by-step implementation of the system.

1) Environment Setup : Python Installation and Configuration: Through using Google Colab for deep-learning model training, the development environment was designed.

2) Installation of Required Libraries : To ensure smooth running of the project the following basic libraries got installed.

a) TensorFlow and Keras: For the construction and the training of the deep learning model.

b) OpenCV: For the preprocessing, for the character segmentation which is very important and, finally, the real-time image handling.

c) NumPy and Pandas: For the numerical operations as well as dataset handling.

d) Matplotlib and Seaborn: For the visualization of the training progress and evaluation metrics.

3) Structuring the Project : It was a large project, which was mainly about the use of structured directories that was established for each task's folder.

a) Dataset Storage: A folder in which there were the Telugu handwritten character images were kept.

b) Preprocessing Scripts: The scripts which included noise removal, grayscale conversion, and binarization were stored separately.

c) Model Training and Evaluation: The training of deep learning models and the evaluation of the response under scripts that were specifically designed for each case thus, leading to more scalable and reusable code.

4) Data Processing and Augmentation :

a) Dataset Acquisition: The setup included the usage of a hand operated text Telugu dataset which included individual characters and words handwritten.

b) Preprocessing Techniques: Every image, in turn, was resized to 64×64 pixels of equally the same scale for correct input dimensions.

c) Image Enhancement: Grayscale conversion and binarization (Otsu's method) to increase contrast.

d) Data Augmentation: In order to make the model more unbreakable, the additional training samples were created by the following random rotations , scaling, and noise reduction.

5) Model Architecture and Training Configuration: A combination of hybrid deep learning model using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) was invented for very accurate computational vision of sequences.

a) Feature Extraction using CNN: 1.The convolutional layers had three of them with a ReLU activation to bring out special features in space. 2.Max-pooling layers for the main patterns. 3.Dropout layers for reducing overfitting.

b) Sequence Learning using LSTMs: 1.Bidirectional Long Short-Term Memory (BiLSTM) are the ones that could be used to go both ways in to the deep to understand the sequential connections and requirements in handwritten words. 2.Dropout used to prevent overfitting.

c) Training Configuration: 1.Adam with an adaptive learning rate is the most effective Optimizer. 2.Loss Function: Connectionist Temporal Classification (CTC) for the alignment-free sequence learning. 3.Batch Size: 32 4.Epochs: 50–100

6) Model Training and Performance Analysis : 1.They preprocessed the model on the provided dataset with the hyperparameters that were optimized. 2.Due to several repetitions of training, the highest test accuracy achieved was 82.743.Confusion Matrix Analysis brought out the maximum number of mistakes in terms of mixing characters with one another.

7) Performance Evaluation : Tested the model model on unseen data. Produced performance metrics like a confusion matrix and classification report to ensure accuracy of the model.

8) Verification: Running the system on handwritten test data and analyzing the pairs of predicted and actual text was carried out.

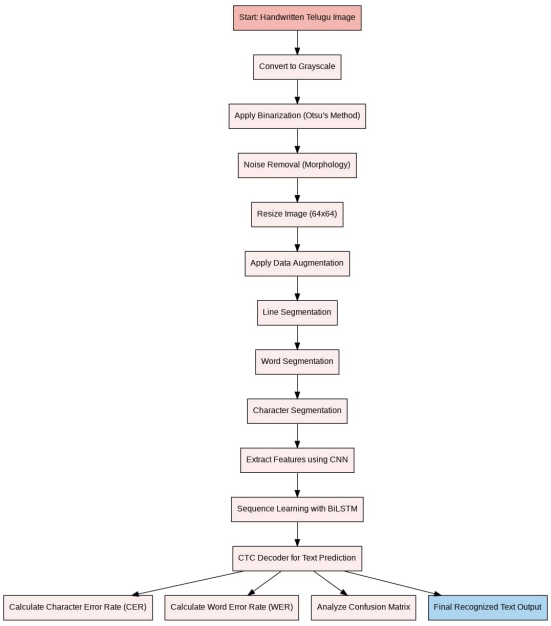

Fig. 2. Process Flow Diagram

B. Results

a) Model Performance Evaluation: The suggested Multilingual Handwritten Character recognition system has been done using various performance metrics which include Accuracy,Precision,Recall,F1-score.The system accurately identifies Telugu handwritten text using Convolutional Neural Networks (CNNs) for feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) for the learning of sequence and at last using CTC decoding for final predictions.

b) Confusion Matrix Analysis: By using confusion matrix we can analyze the most frequently missclassified characters such as visually similar characters,rare occurerence of characters and connection of strokes within the characters which led to decrease in accuracy.

c) Precision-Recall Curve: The Precision-Recall Curve as shown in the Fig. 6. helps us to handle the imbalanced

Metric	Value
Accuracy	82.94%
Precision	73.92%
Recall	71.91%
F1-score	71.96%

Fig. 3. Performance Metrics

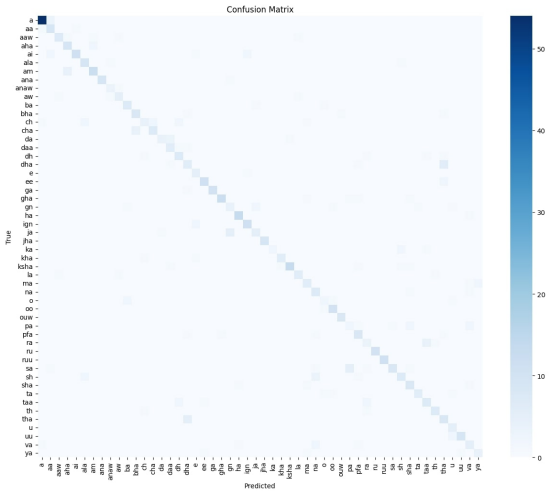

Fig. 4. Confusion Matrix

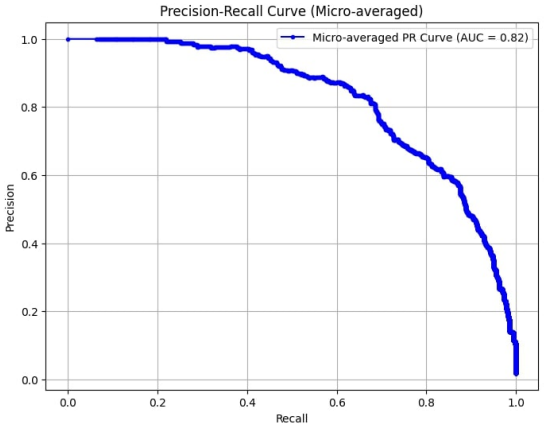

Fig. 5. Precision-Recall Curve

data.It helps us to distinguish correct and false classifications.

The Fig. 6. shows the output which provides the prediction given by the model when user inserts an image for character recognition.

The below drawn TABLE 1 is a observation which gives us the Precision,Recall,F1-score,Support for the 34 classes which we used in our model training.

Precision measures that how many of the predicted positive cases were correct actually and recall tells us that how many actual positive cases were found correctly. The F1-score balances the recall and precision and gives us the actual measure of the performance.

Class	Precision	Recall	F1-score	Support
0	0.947	0.947	0.947	57
1	0.600	0.818	0.692	11
2	0.778	0.636	0.700	11
3	0.643	0.818	0.720	11
4	0.917	0.688	0.786	16
5	0.692	0.818	0.750	11
6	0.800	0.750	0.774	16
7	1.000	0.900	0.947	10
8	0.600	0.750	0.667	4
9	0.625	0.625	0.625	8
10	0.700	0.778	0.737	9
11	0.667	0.727	0.696	11
12	0.571	0.333	0.421	12
13	0.778	0.636	0.700	11
14	0.800	0.500	0.615	8
15	0.545	0.750	0.632	8
16	0.583	0.636	0.609	11
17	0.400	0.429	0.414	14
18	0.556	0.833	0.667	6
19	0.846	0.846	0.846	13
20	1.000	0.909	0.952	11
21	0.923	0.750	0.828	16
22	0.400	0.400	0.400	10
23	0.867	0.929	0.897	14
24	0.786	0.846	0.815	13
25	0.625	0.455	0.526	11
26	1.000	1.000	1.000	9
27	1.000	0.400	0.571	5
28	0.857	0.750	0.800	8
29	0.929	0.722	0.813	18
30	0.750	0.667	0.706	9
31	0.600	0.667	0.632	9
32	0.500	0.778	0.609	9
33	0.400	0.333	0.364	6
34	0.769	0.909	0.833	11
Accuracy		0.719		
Macro Avg	0.707	0.698	0.691	591
Weighted Avg	0.739	0.719	0.720	591

TABLE I
PERFORMANCE METRICS PER CLASS AND AVERAGE METRICS.

V. CONCLUSION AND FUTURE SCOPE

In conclusion, The Multilingual Character Recognition has heped a lot in the improvement optical character recognition (OCR) of the most complex scripts .The system is successful in recognizing Telugu handwritten characters by combining Convolutional Neural Networks (CNNs) for feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) networks for sequence recognition.Preprocessing played a crucial role in increasing accuracy of the system. The model obtained

Fig. 6. Real Time Prediction for user provided image

an accuracy of 82.74,by using performance metrics such as Precision,Recall,Accuracy,F1-score which helped in providing a depth analysis on systems performance.Moreover this model is capable of recognizing various handwritings and fonts. The usage of confusion matrix has helped to analyze the most frequent misclassifications of characters which improved the systems accuracy. The future advancement can be increasing the size of the dataset with various handwritings,better segmentation algorithms for clearly separating the texts. Another future scope can be the improvement in the error fixing procedures such as using Hybrid AI which is integrating OCR with NLP which will lead to optimizing the accuracy in both the systems. This model can be further used in the mobile and web applications ensuring that it can be used for large number of textual formats and according to the preference of the individuals.

REFERENCES

[1] BNV Buddaraju Revathi, Ajay Dilip Kumar Marapatla, Kagitha Veeramanikanta, Katta Dinesh, and Maddirala Supraja. Optical character recognition for telugu handwritten text using squeezenet convolutional neural networks model. *International Journal of Advances in Applied Sciences*, 13(3), 2024.

[2] M Swamy Das, CRK Reddy, K Rahul, and A Govardhan. Multilingual optical character recognition system for printed english and telugu base characters. *International Journal of Science and Advanced Technology (ISSN 2221-8386)*, 1(4):106–111, 2011.

[3] Ashlin Deepa, Guru Sai Jayanth Kalluri, Zeeshan Mohammed, Mantri Pramod Sai Sushank, Atul Negi, et al. A novel approach to recognize handwritten telugu words using character level cnn. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE, 2023.

[4] Srinivasa Rao Dhanikonda, Ponnuru Sowjanya, M Laxmidevi Ramanaiah, Rahul Joshi, BH Krishna Mohan, Dharmesh Dhabliya, and N Kannaiya Raja. An efficient deep learning model with interrelated tagging prototype with segmentation for telugu optical character recognition. *Scientific Programming*, 2022(1):1059004, 2022.

[5] Tejasree Ganji, Muni Sekhar Velpuru, and Raman Dugyala. Multi variant handwritten telugu character recognition using transfer learning. In *IOP Conference Series: Materials Science and Engineering*, volume 1042, page 012026. IOP Publishing, 2021.

[6] John Jomy, KV Pramod, and Balakrishnan Kannan. Handwritten character recognition of south indian scripts: a review. *arXiv preprint arXiv:1106.0107*, 2011.

[7] P Pavan Kumar, Chakravarthy Bhagvati, Atul Negi, Arun Agarwal, and Bulusu Lakshmana Deekshatulu. Towards improving the accuracy of telugu ocr systems. In *2011 International Conference on Document Analysis and Recognition*, pages 910–914. IEEE, 2011.

[8] SB Madhu, CV Aravinda, and MS Sannidhan. Handwritten kannada character recognition using convolutional neural networks and transfer learning. In *Journal of Physics: Conference Series*, volume 2571, page 012012. IOP Publishing, 2023.

[9] B Meena, K VENKATA RAO, and SURESH CHITTINENI. A novel method to auto configure convolution neural network model using soft computing technique to recognize telugu hand-written character for better accuracy. *Journal of Theoretical and Applied Information Technology*, 100(18), 2022.

[10] Konkimalla Chandra Prakash, YM Srikar, Gayam Trishal, Souraj Mandal, and Sumohana S Channappayya. Optical character recognition (ocr) for telugu: Database, algorithm and application. In *2018 25Th IEEE international conference on image processing (ICIP)*, pages 3963–3967. IEEE, 2018.

[11] SNS Rajasekaran and BL Deekshatulu. Recognition of printed telugu characters. *Computer graphics and image processing*, 6(4):335–360, 1977.

[12] C Sunitha Ram, UTS Abhiroopika, and S Chinmayee. Handwritten telugu compound character prediction using convolutional neural network. 2023.

[13] N Shobha Rani and T Vasudev. Automatic detection of telugu single and multi-character text blocks in handwritten words. In *2015 International Conference on Computing and Network Communications (CoCoNet)*, pages 234–240. IEEE, 2015.

[14] BV Subba Rao, J Nageswara Rao2 Bandi Vamsi Venkata, and Nagaraju Thatha4 Katta Subba Rao. A unicode based deep handwritten character recognition model for telugu to english language translation. *IJCSNS*, 24(2):101, 2024.

[15] Vempati Lakshmi Sravani and Piyush Pratap Singh. Hybrid approach for recognition of isolated handwritten fraction notations in telugu script. *International Journal of Computer Applications*, 975:8887.