International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 17s, 2025
https://theaspd.com/index.php

Enhanced Augmented CNN Model Using Optimized Optical Character Recognition For Off-Line Telugu Characters

Dumpal Koteswararao[1], Dr. Nagaratna P Hedge[2]
[1]Research Scholar, Department of CSE, Osmania University.
[2]Professor, Department of CSE, Vasavi College of Engineering
[1]dumpal.koteswararao@gmail.com, [2]nagaratnaph@staff.vce.ac.in

Abstract:
The Indian Constitution acknowledges the significance of Telugu, Tamil, Malayalam, and Kannada as languages. Worldwide, over 90 million people speak Telugu, a language from South India. Digitalizing books and unstructured documents are one of the applications for Telugu optical character recognition (OCR), which improves human-to-human communication. The training of optical character recognition systems is more extensive for international languages such as English and German than for regional languages such as Telugu, Tamil, Malayalam, etc. The main difficulty in developing TOCR is the large number of unique characters used in the Telugu language. This paper makes two contributions to help with this: (i) a collection of Telugu characters written by hand, and (ii) a convolutional neural network (CNN) model with enhancements to identify the hand-written characters in the scanned text.
Keywords: *Enhanced Convolutional Neural Network (ECNN), Segmentation, TOCR, Telugu character recognition, Telugu Language, Training*

INTRODUCTION

The states of Andhra Pradesh and Telangana have Telugu as their primary language of communication. According to the counts in the Ethnologue enumeration of languages by native speaker count, Telugu ranks the fourth most spoken language of the Indian subcontinent and is placed 15th in the list of the world's most spoken languages [1]. OCR is the process by which a computer extracts text from an image and the result is a computer-readable text.In actuality, a vast number of imaged articles circulate throughout the digital domain. It's time to convert those scanned papers into editable format. TOCR is a methodology that enables the recognition and digitization of Telugu text from scanned images or handwritten notes. This technology has numerous applications in the fields of education, government, publishing, and archiving, among others. Telugu character shapes are numerous, including simple and compound symbols composed of 16-vowels (referred to as achus) and 36-consonants (designated as hallus). On top of these modifiers can be applied, which causes complexity in recognizing the characters, making it a challenging task for the OCR systems to accurately recognize and digitize the text. To address this challenge, researchers and developers have used various techniques such as feature extraction, neural networks, and machine learning algorithms to enhance the accuracy of TOCR method. In feature extraction methods, mathematical formulas are applied to find the important characteristics of the characters, and in the neural and machine learning methods, the patterns and structures of the characters are learned by using sample data.The most potential way in the field of Telugu OCR technology is deep learning technique techniques like CNNs. image recognition is used for telugu ocr systems because Convolutional Neural Networks (CNNs) are known to perform well and give high accuracy for image recognition problems.Telugu OCR is a rapidly evolving field with numerous applications and challenges. With the advancements in deep learning and computer vision technologies, the accuracy and efficiency of Telugu OCR systems are expected to continue to improve in the coming years.
Neural networks and machine learning algorithms have also been used to address the challenges faced by OCR systems in Telugu script. These techniques use training data to learn the patterns and structures of the characters, enabling the OCR system to accurately recognize the characters.
Deep Learning Techniques in Telugu OCR: One of the most promising approaches in Telugu OCR technology is the use of deep learning techniques, such as Convolutional Neural Networks (CNNs). CNNs

are particularly well-suited for image recognition tasks and have been used to achieve high accuracy in Telugu OCR systems.

The two key steps of OCR for producing high-level accuracy in character recognition are segmentation and classification.

In this paper, the following contributions are addressed:

i) Introduced a new dataset for Telugu characters with 52 cateogies (16 vowels and 36 consonants) and 25068 samples.

ii) Proposed an enhanced CNN model to recognize 16 vowels (called achus) and 36 consonants (called hallus) with 99.37% accuracy.

The rest of the paper is organized as follows:

i) Section 2 discussed about the literature works on Telugu OCR related to off-line characters.

ii) Section 3 talks about te methodology.

a) Subsection 3.1 discussed about the new dataset proposed.

b) 3.2 discussed about the enhanced CNN framework to recognize off-line characters.

iii)In section 4, the results are disussed. iv) Conclusion and future works are discussed in section 5.

2 LITERATURE WORK

One of the most researched challenges in pattern recognition is optical character recognition (OCR). Until recently, the dominating approach was featuring engineering, which included features such as Wavelet features, Gabor features, Circular features, Skeleton features, and so on; [2],[3],[4], followed by one of the leading classification techniques. We decided to employ CNNs for Telugu character recognition because of their recent and remarkable achievement in feature learning. Atul Negi and Samit Kumar Pradhan, [5] proposed for the hand-written basic Telugu characters. String comparison for character recognition and tried data structure to improve the time complexity are discussed. Due to the approximate string matching the high accuracy may not be guaranteed here. Vijaya Krishna Sonthi, et.al, [6] proposed an innovative multi-objective mayfly optimization with deep learning MOMFO-DL model for the recognition of hand-written Telugu characters and reached 99% accuracy level. But this method also works on exact matching.

Srilakshmi Inuganti, et.al, [7] presented "aimed to recognize strokes in the characters and achieved 93% accuracy. Tejasree Ganji, et.al, [8] presented applied VGG-16 for recognizing characters and achieved 92% accuracy over 1600 sample characters. Comparison study was discussed by Muni Sekhar Velpuru, et.al, [9], where all the available OCR models are compared. N Prameela, et.al [10], proposed 47 categories and 200 samples are taken to train the model and achieved 87.6% accuracy, which can be improved in further works. Ananda Kumar Kinjarapu, et. al [11] achieved 99% accuracy by considering 28-sample features only. Optimizer Model using Deep-Learning was discussed by Vijaya Krishna Sonthi, et. al, [12], a new technique DLTCR-PHWC derived to detect printed and hand-written characters in the scanned images and achieved 99.22% accuracy. Panyam Narahari Sastry, et. al, [13], works with Zoning Features using Nearest Neighborhood Classifier (NNC) method is used for identifying and classifying, achieved 78% accuracy only. Recently Ashlin Deepa R. N, et. al, [14], achieved 72.6% accuracy overall for all distinct characters which can be improved by training the model with huge dataset.

3 Proposed Methodology:

The pipeline followed for OCR system is: scanning documents – preprocessing – segmentation – classification of characters. Our paper introduces novelties in the dataset and classifier.

3.1 Dataset

Due to lack of proper dataset for hand-written Telugu characters, many OCR systems are not able to give proper results. We, consider this issue and trying to supply one basic dataset, which can be useful for all the researchers. Entire dataset creation was partitioned into two parts, part-1 is useful for recognizing 52 main characters including vowels and consonants. Part-2 for modifiers. In this paper, part-1 dataset is available and respective classification model is discussed in the next step.

25068 samples are considered for 52-classes training set with approximately 500 images per class, and 4654 samples are taken for testing purpose. All these samples are taken from different people including augmentation, so it was a unique dataset with each image size is 32 X 32.

Table 1, shows all the Vowels and Consonants of off-line Telugu characters with class labels and samples taken for each label. Each label is mapped with respective Unicode values and each Unicode values represents respective Telugu characters, Which can be check using Hyper Text Markup Language (HTML).

Table 1: Vowels and Consonants of off-line Telugu Characters with labels

Class	Image	Unicode	Count	Class	Image	Unicode	Count
1		&#3077;	497	27		&#3103;	291
2		&#3078;	497	28		&#3104;	497
3		&#3079;	499	29		&#3105;	490
4		&#3080;	498	30		&#3106;	489
5		&#3081;	496	31		&#3107;	488
6		&#3082;	494	32		&#3108;	487
7		&#3083;	492	33		&#3109;	493
8		&#3168;	368	34		&#3110;	490
9		&#3086;	494	35		&#3111;	492
10		&#3087;	495	36		&#3112;	494
11		&#3088;	484	37		&#3114;	490
12		&#3090;	480	38		&#3114;	490
13		&#3091;	489	39		&#3116;	487
14		&#3092;	494	40		&#3117;	495
15		అం	494	41		&#3118;	477
16		అః	497	42		&#3119;	493
17		&#3093;	495	43		&#3120;	493
18		&#3094;	494	44		&#3122;	496
19		&#3095;	494	45		&#3125;	492
20		&#3096;	497	46		&#3126;	490
21		&#3097;	493	47		&#3126;	492

22		&#3098;	493	48		&#3128;	488
23		&#3099;	493	49		&#3129;	492
24		&#3100;	390	50		&#3123;	493
25		&#3101;	389	51		క్శ	494
26		&#3102;	498	52		&#3121;	492

3.2 **Classifier**

The performance of the classifier of an OCR system plays an important role in its efficiency. The class for each of the classifiers will be predicted from the character level segmentations.

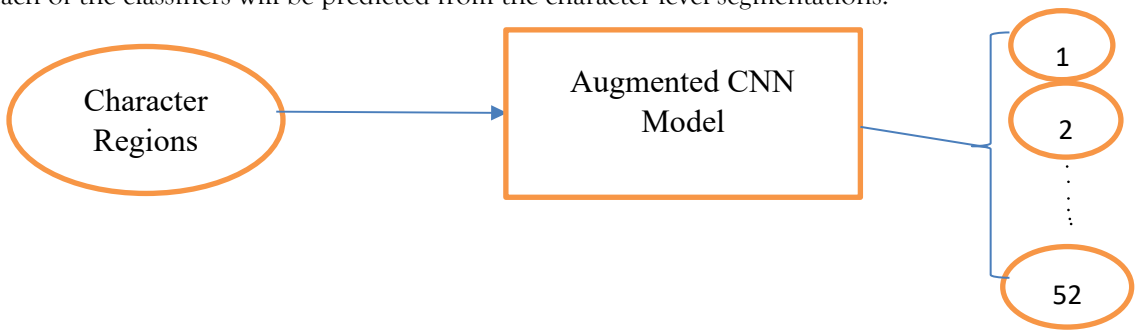


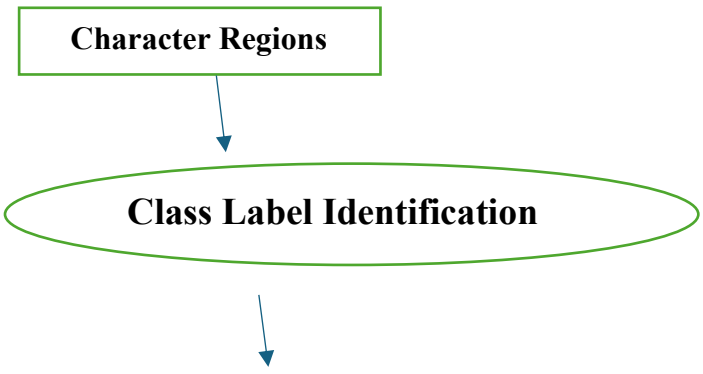
Figure 1: Character regions to class labels

Figure 1, gives a simple flow of our work, where character regions are given as input to the augemented CNN model which gives a respective class label. In this research work, we explored CNNs on categorising the characters and proposed a new architecture for it, both inspired by the fact that deep neural networks have been very successful in learning features.

A CNN is a special type of feed-forward neural network, or a network with a number of layers, based off of biological processes. Each Telugu letters consists of two elements; the first element prints the main character and the second element the modifiers. In this paper, We first proposed a CNN model for the first component.

Augmented CNN model comprises the following steps:

1) 13 Convolutions and 2 Fully-connected layers.
2) Model training by considering training dataset.
3) Finding accuracy based on testing dataset.



Conv1 (3X3)
Maxpooling 1
Conv2(3X3)
Maxpooling 2
Conv3 (3X3)
Maxpooling 3
Conv4 (3X3)
Maxpooling 4
Conv5 (3X3)
Maxpooling 5
Conv6 (3X3)
Maxpooling 6
Conv7 (3X3)
Maxpooling 7
Conv8 (3X3)
Maxpooling 8
Conv9 (3X3)
Maxpooling 9
Conv10 (3X3)
Maxpooling 10
Conv11 (3X3)
Maxpooling 11
Conv12 (3X3)
Maxpooling 12
Conv13 (3X3)
Maxpooling 13
Fc-1
Fc-2

Figure 2: Telugu Han-written Main Character CNN (THMC-CNN) model to recognize offline characters

RESULTS AND DISCUSSION:

In this article we discuss the details of our experiment and the results we obtained with our proposed approach. As mentioned before, 25068 samples are used for training the network. The performance was verified with 4654 samples. We set a batch size of 50 according to the size of the training data. At first, we trained our network with a Gradient Descent optimizer. The performance was only 90% after 400–500 epochs, which is not even enough. We could achieve significantly higher accuracy by using the Adam optimizer with 100 epochs. If no improvement is achieved in the validation accuracy within 10 epochs, we terminate the training process. We trained our model on Google Colab with 48 GB of RAM.

We would like to remark that commonly used CNN structures [19, 20] of Cifar and VGG- 16 had also been trained by the same approach beside the CNN structure in Fig. 2. As further described below, this was largely performed for comparison purposes.

Table 2: CNN Accuracies for Main Character Classification

Network	Accuracy (%)
Cifar	94.45
VGG-16	98.65
THMC-CNN	99.37

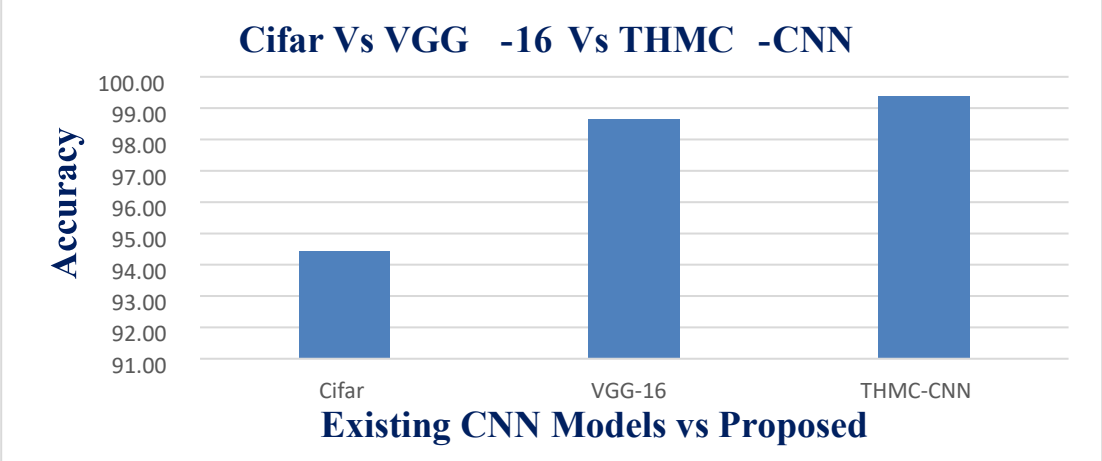


Figure 3: CNN Accuracies for Main Character Classification

CONCLUSION:

In conclusion, the development of OCR systems for the Telugu script is a rapidly evolving field with numerous challenges and opportunities. With the advancements in deep learning and computer vision technologies, the accuracy and efficiency of Telugu OCR systems are expected to continue to improve in the coming years. The development of efficient and accurate OCR systems for Telugu script will play a critical role in digitizing and preserving the rich cultural and literary heritage of the Telugu language. We demonstrated a Telugu OCR solution that included a database for main characters component and a classifier. The proposed work can be extended in two categories; one is creating and including modifiers dataset to the main characters component dataset and the other one is training one more CNN model to recognize modifiers involved in the character segmentations.

Declarations

- Funding available - No
- Conflict of interests/ Competing interests - No
- Ethics approval- yes

REFERENCES

[1]Wikipedia contributors, "Telugu language — Wikipedia, the free encyclopedia," 2018, [Online; accessed 13- February-2018].

[2]Arja Rajesh Babu, "Ocr for printed telugu documents," Diss. Indian Institute of Technology Bombay Mumbai, 2014.

[3]Ramanathan, Arun S Nair, L Thaneshwaran, S Ponmathavan, N Valliappan, and KP Soman, "Robust feature extraction technique for optical character recognition," in Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on. IEEE, 2009, pp. 573–575.

[4]Peifeng Hu, Yannan Zhao, Zehong Yang, and Jiaqin Wang, "Recognition of gray character using gabor filters," in Information Fusion, 2002. Proceedings of the Fifth International Conference on. IEEE, 2002, vol. 1, pp. 419–424

[5]Samit Kumar Pradhan and Atul Negi. 2012. A syntactic PR approach to Telugu handwritten character recognition. In Proceeding of the workshop on Document Analysis and Recognition (DAR '12). Association for Computing Machinery, New York, NY, USA, 147–153.

[6]Vijaya Krishna Sonthi, S. Nagarajan, and N. Krishnaraj. 2022. An Intelligent Telugu Handwritten Character Recognition using Multi-Objective Mayfly Optimization with Deep Learning Based DenseNet Model. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (March 2022). https://doi.org/10.1145/3520439

[7]Srilakshmi Inuganti and Rajeshwara Rao Ramisetty, 2021, Online Handwritten Telugu Character Recognition Using Normalized Differential Chain Code Feature, Computing Technology and Information Management, Volume 18, Special Issue, pp 1-15.

[8]Tejasree Ganji et al 2021 "Multi Variant Handwritten Telugu Character Recognition Using Transfer Learning" IOP Conf. Ser.: Mater. Sci. Eng. 1042 012026

[9]Muni Sekhar Velpuru, Priyadarshini Chatterjee, G Tejasree, M Ravi Kumar, S Nageswara Rao "Comprehensive study of Deep learning-based Telugu OCR", IEEE ICSSIT 2020, 22 August 2020.

[10] Prameela, N & Pimpalshende, Anjusha & Karthik, R. (2017). Off-line Telugu handwritten characters recognition using optical character recognition. 223-226. 10.1109/ICECA.2017.8212801.

[11] Kinjarapu, Ananda & Yelavarti, Kalyan & Valurouthu, Kamakshi. (2016). Online recognition of handwritten Telugu script characters. 426-432. 10.1109/SCOPES.2016.7955866.

[12] Krishna, Sonthi & Nagarajan, S. & Krishnaraj, N.. (2021). Automated Telugu Printed and Handwritten Character Recognition in Single Image using Aquila Optimizer based Deep Learning Model. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0121275.

[13] Narahari, Panyam & T.R., Vijaya & Rao, N V & Rajinikanth, T.V. & Wahab, Abdul. (2015). Telugu Handwritten Character Recognition Using Zoning Features. 2014 International Conference on IT Convergence and Security, ICITCS 2014. 10.1109/ICITCS.2014.7021817.

[14] Ashlin, Deepa & Vijayalata, Y. & Negi, Atul. (2022). Document Text Analysis and Recognition of Handwritten Telugu Scripts. 462-466. 10.1109/ICCCMLA56841.2022.9989012.