Multi Variant HandWritten Telugu Character Recognition Using Transfer Learning

Tejasree Ganji 1, Muni Sekhar Velpuru2, Raman Dugala3,  
1 Department of Computer science and Engineering,  
2 Department of Information Technology, Vardhaman College of Engineering and Technology  
*tejasreeganji@gmail.com*

Abstract. Optical Character Recognition (OCR) has become one of the most important techniques in computer vision, given that it can easily obtain information from various images. However, existing OCR techniques cannot recognition Telugu literature characters (Handwritten Golusu Kattu writing) due to a lack of datasets and trained deep Convolutional Neural Networks (CNN). Since the Kakatiya Empire (12th to 14th century) the glorious era of Telugu literature spread across the region. Thereupon, several handwritten documents consist of ancient knowledge, health care tips, wealth information, and several land records written in Telugu Golusukattu writing. Therefore, getting that information has become a major problem because of a lack of expertise in Golusukattu writing skills in skills. In order to solve the above problem, we are proposing deep learning aided- OCR for Telugu literature.

Keywords: Optical Character Recognition, Deep Convolutional Neural Network, Telugu handwritten recognition, Deep Learning.

1. Introduction

In this era, there is an increasing demand for applications that can recognize handwritten characters automatically when it is fed to the computer. Now a day there is an immense demand for storing this kind of information available in either books or in any kind of documents into a computer for future use which can be retrieved later by search technique. This can be done by scanning the information which had to store and save them. But the problem with this is one can face difficulty to read the document as we cannot search the content of the image. The reason for this difficulty is those handwritten font’s characters differ from the font of the characters in computer systems. Thus, the computer fails to recognize the characters while reading them.

This concept of storing the contents of paper documents in a computer storage place and then reading and searching the content is called document processing. Sometimes in this document processing, we need to process the information that is related to languages other than the English in the world. For this document processing, we need a software system called the character recognition system. This process is also called Document Image Analysis (DIA). Thus, we aim to build a character recognition system that can perform DIA which automatically transforms input documents into electronic format. There are distinct techniques to carry out this process, among these techniques we had chosen the Optical Character Recognition technique.

The goal of Optical Character Recognition (OCR) is used to classify optical patterns into corresponding alphanumeric or other characters. It is a type of document image analysis where a scanned digital image contains a machine-printed document or handwritten script is input to an OCR software engine and translate it into an editable machine-readable digital text format (like ASCII text).

Optical character recognition is the most important technique in computer vision. It can easily obtain information from various images, but it cannot recognize Telugu handwritten characters due to a lack of data sets and trained deep convolutional neural networks. Optical character recognition is the process of converting scanned images of handwritten, typed, or printed text into machine-encoded text. OCR technology deals, with the problem of recognizing all kinds of different characters. Handwritten and printed characters can be recognized and converted into a machine-readable or digital data format. By using OCR we can transform serial numbers or code consisting of numbers and letters into a digital output.
In our research work, we had focused on character image classification, segmentation, and recognition. We had implemented the proposed model based on improved deep learning approaches for Telugu Handwritten Character (scanned Handwritten image) segmentation, classification, and detection. A comparison of the proposed model along with other models demonstrates greater performance against similar. The main contribution of this research work is summarized as follows:

1. The proposed model is applied to classify Handwritten Characters.
2. The proposed Telugu Character Recognition transfer learning model “TCR_new_model.h5” is designed from the VGGNET model.
3. For the models, from encoding to decoding unit concatenation is applied to the feature maps.
4. The proposed model provides a good outcome compared to other models.

The paper is organized as follows: Section 2 discusses related work. Section 3 describes the methodology and proposed architecture, section 4 explains the dataset and experimental results. Section 5 describes the conclusion and future work.

2. Related Work

A global survey has been done on Optical Character Recognition, Segmentation, and Deep Learning approaches where various learning techniques, models, and recently proposed training approaches are discussed.

Atul Negi et al. [1], implemented a model using zoning and structural feature vector to recognize the isolated text and they define cavities as a new set of structural features for recognition. In the segmentation process, they gave input as a document image, and that image is analyzed in 2 phases—top-down analysis and bottom-up analysis. The proposed method achieved an accuracy of 97%-98%.

Muni Sekhar Velpuru et al.[4], they surveyed Optical Character Recognition, Segmentation, and Deep learning. Optical Character Recognition (OCR) is used to recognize text in the images. Images mean scanned documents or pictures. By using OCR, there is no need to digitize the image text. Another advanced technology that is extensively used in OCR is deep earning. Segmentation is used to segment the words into characters.

K. Mohana Lakshmi et al. [10], developed an efficient algorithm to identify individual handwritten Telugu characters based on HOG (histogram of oriented gradient) features and handwritten Telugu text efficiently. Experimental results are analyzed using the Bayesian network algorithm. Depending upon the true positives and true negatives accuracy has been calculated. They achieved the recognition rate for the Telugu script is 87.5%.

C. Vikram et al. [6], proposed a model for the recognition of Telugu handwritten characters, they approach the pre-processing and recognition. The pre-processing is based on ANN (Artificial Neural Network) and MLP (multi-layer perceptron). The recognition is based on hybrid optical HMM/ANN (Hidden Markov model/Artificial Neural Network) and the Multi-Layer Perceptron model applies to handwritten Telugu scripts. For recognition of unconstrained offline handwritten text lines, they used a hybrid HMM/ANN (Hidden Markov Model/Artificial Neural Network) system that is proposed by
Salvador Espanaet [1]. The Proposed scheme has a recognition accuracy of 85%.

Panyam Narahari Sastry et al. [8], In this paper, zonal based feature extraction is used for feature extraction, the character image is divided into a predefined number of zones and a statistical feature is computed from each of these zones. The features of all the zones in the image from a feature vector which is used for handwritten character recognition. In this, they have used NNC (nearest neighbor classifier) i.e., Euclidean distance for identifying and classification for Telugu text. Using the zoning method the recognition accuracy found to be 78%.

K.Vijay Kumar et al. [11] proposed online handwritten character recognition for the Telugu language using support vector machines. To interpret characters automatically automated character recognition is used and it is useful while processing a large number of handwritten characters. To overcome the variations in font and size of Handwritten Character Recognition they used the following techniques. 1) to review the issues and techniques for HCR which are: i) Preprocessing ii) Step stroke Pre-classification iii) Feature Extraction, and 2) to review the support vector machine (SVM) as the classification method. The proposed model achieved an accuracy of around 82.96%.

Konkumalla Chandra Prakash et al. [9]; developed the dataset and classifier also introduces preprocessing, segmentation (word segmentation, character segmentation), recognition, and convolution neural network (CNN). The performance of the OCR system depends on its classifier. They have used SVM based classifiers for character classification. They have used 2CNN architecture for classifying the characters. They have developed an android app for this project, which converts an online image to text with industry-standard; it works on any android device. It uses a camera or gallery for images and it is publicly available. The experimental results show a better accuracy of about 98.60%.

SD Prasad et al. [15]; to improve the accuracy in Telugu Character recognition they used Zonal-based Feature Extraction. In Zonal based Feature Extraction, they presented two methods: 1) based on Genetic Algorithm and uses Adaptive Zoning topology with extracted geometric features. 2) Zoning is done in a static way and uses distance, density-based features. For feature extraction, they use a zone-based approach. And also, they use 3 types of features. They are distance, density, and geometric based features.

C.Vasantha Laxmi et al. [2]; proposed A high accuracy OCR System for Printed Telugu Text. In Pre-processing technique grayscale image is converted to a binary image, neural recognizers are used for identification of basic symbols which founds the fundamental units for segmentation purpose. The main feature of the proposed model is that it can handle multiple sizes and multiple fonts. That outcome of it can be directly implemented to other Indian languages software which supports translation facility into Telugu.

A Ashiqzzaman et al. [12]; the author has proposed deep learning neural networks using appropriate activation function and regularization layer. Developed a method that can increase the accuracy of this process by using a convolution neural network (CNN) in place of MLP. CMATERDB Arabic handwritten digit dataset is used to evaluate experimental results. This proposed model will give 97.4% accuracy.

R Vaidya et al. [13]; Author has proposed an innovative method for offline handwritten character detection using deep neural networks. They have designed a segmentation-based handwritten character recognition system for image segmentation. They had used Open CV for performing Image Processing and for training a neural network they used Tensor Flow. To train the neural network they comprised a stack of Convolution layer and MAX-pooling layer followed by a fully connected layer.

Meiyin Wu et al. [14]; In this, they applied deep learning to real-world handwritten character recognition, and explored the two mainstream algorithms of deep learning: the Convolutional Neural Network (CNN) and the Deep Belief Network (DBN). They conduct the performance evaluation for CNN and DBN on the MNIST database and the real-world handwritten character database. And they
achieved the classification accuracy rate of CNN and DBN on the MNIST database is 99.28% and 98.12%. And on the real-world handwritten character database is 92.91% and 91.66%. They analyzed differences between CNN and DBN by comparing the experimental results.

Table 1. Summary of Handwritten character recognition using different algorithm & methods

| S.NO | Author Name            | Methods Used                             | Dataset                  | Validation accuracy |
|------|------------------------|------------------------------------------|--------------------------|---------------------|
| 1    | Atul Negi et al. [1]   | Zoning and structural Feature Vector     | Telugu character dataset | 92%                 |
| 2    | K. MohanaLakshmi et al. [10] | Histogram of Oriented gradient Features | Telugu word images       | 87.5%.              |
| 3    | C. Vikram et al. [6] | Multi Layer Perceptron And Hidden Markov Model | Telugu Scripts | 85%                |
| 4    | Panyam Narahari Sastry et al. [8] | Zonal Based feature Extraction | Telugu handwritten characters | 78%                |
| 5    | K. Vijay Kumar et al. [11] | Support Vector Machine                    | Own dataset              | 82.96%              |
| 6    | Konkimmalla Chandra Prakash et al [9] | SVM based Classifier                   | Own dataset              | 98.60%              |
| 7    | SD Prasad et al [15] | Zonal based Feature Extraction            | Telugu characters        |                     |
| 8    | C. Vasantha laxmi et al [2] | A High accuracy OCR system                | Basic symbols            |                     |
| 9    | A Ashipuzzaman et al[12] | Deep Learning Neural Network              | CMATERDB Arabic Hand written digit dataset | 97.4%              |
| 10   | R Vaidya et,al[13] | Deep Neural Network                       | MNIST                    |                     |
| 11   | Meiyn Wuet,al[15] | Convolutional Neural Network              | MNIST and Real World hand Written Database | 91.66%             |

3. Methodology
According to the literature, survey researchers had developed many applications in the field of handwritten character recognition for the automatic computer-aided detection system. The main function of handwritten character recognition is classifying and recognizing the handwritten characters. The image of the handwritten text may be sensed “offline” from a piece of paper by optical scanning (Optical Character Recognition) or Intelligent word Recognition. A handwriting recognition system handles formatting, performs correct segmentation into characters, and finds the most plausible words.

The three objectives of the work are:
- Assess the current level of technology in Telugu OCR
- Identify opportunities and challenges facing by Telugu OCR
- Converting ancient Telugu documents into digital records (Converting Telugu handwritten literature into digital records.)

**Image Acquisition:** Handwritten TELUGU Character Dataset includes 1600 characters. It has been divided up into training and testing sets of 70% and 30%. The images are of grayscale. The size of each image is 52x52. In the preprocessing stage image is resized into the following sizes: 224x224. 128x128,
64x64 and 52x52 from the dimensions 52x52. Each image is normalized as shown in figure 3. Shows one scanned character sample image on the left side and its corresponding recognition map with the ground truth on the right side.

A. Data Collection: This dataset consists of 1400 unique characters, 52 sets of each character handwritten images written by 52 distinct writers. In addition, 12 people under 18 years, 7 people above 45, and the remaining are engineering graduates to enable different handwriting styles. It has the following advantages compared to the offline handwritten Telugu characters.

- Segmentation of characters not required. Because each character scanned as a different image with a good resolution.
- No salt-and-pepper noise, no chance for eroded margins, no skewed while scanning
- No need to spend time on character extraction, only focus on character recognition.

This dataset covers most of all Telugu characters in the literature approximately 1600 characters. It is the largest dataset available for Telugu characters as of now. Moreover, each character image is multiplicities with 12 using Image Data Argumentation (shifting, rotation, and sheering effect).

B. Data Analysis: in this, we created a dataset from collected data. And we collected data from various people. We divided the data as a testing dataset and training dataset. We train the dataset using VGGNET-16.

C. Validation: we cross-check the results from the test dataset.

Architecture: by using the transfer learning algorithm we had proposed a Visual Graphics Group-16 model. The VGGNET can be pre-trained so that it works as a deep feature extractor. CNN's are data-driven learning, helps in higher feature extractors, highly representative but they need sufficient training data. The proposed model has three models. These three models are considered as they had performed well on Telugu Handwritten Character recognition tasks. The proposed Telugu Character Recognition transfer learning model “TCR_new_model.h5” is designed from the VGGNET model with different input sizes.

| Model No. | Input Sizes |
|-----------|-------------|
| Model 1   | (224, 224, 3) |
| Model 2   | (128, 128, 3) |
| Model 3   | (64, 64, 3) |
| Model 4   | (52, 52, 3) |
Table 3. input size and parameters of CNN Model

| Model No. | Input Size | Output Total Parameter | Trainable Parameter | Non Trainable Parameters |
|-----------|------------|------------------------|---------------------|-------------------------|
| Model 1   | (224, 224, 3) | 15,116,112             | 401,424             | 14,714,688              |
| Model 2   | (128, 128, 3) | 14,845,776             | 131,088             | 14,714,688              |
| Model 3   | (64, 64, 3)   | 14,734,586             | 30,456              | 14,714,688              |
| Model 4   | (52, 52, 3)   | 14,722,896             | 8,208               | 14,714,688              |

Configuration of VGG-16:

VGGNET-16 architecture contains six layers and one input layer and one output layer as shown in figure (3). The first layer of architecture contains two convolutional layers and one max pooling layer. The second layer has two convolutional layers and one max pooling layer. The third layer contains three convolutional layers and one max pooling. The fourth layer has three convolutional layers and one max pooling. The fifth layer contains three convolutional layers and one max pooling layer and the last layer contains three dense layers.

Let’s consider input to the network an image of dimensions (224,224,3) as shown in table (2). And the first 2 layers have 64 channels of 3*3 filter size and the same padding. After that, a max pool layer of
stride(2,2), the 2 layers which have convolution layers of 256 filter size and the filter size is (3*3). This process is followed by a max pool layer and a max-pooling layer is the same as the previous layer. There are two sets of three convolutional layers and a max pool layer. Each has 512 filters of (3*3) instead of 11*11 in AlexNet and 7*7 in ZF-Net. But in some of the layers, also uses 1*1 pixel which is used to manipulate the number of input channels. And also there is a padding of 1-pixel done after each convolution layer to prevent the spatial feature of the image.

The proposed VGGNET-16 model is the building block of the convolutional units and architecture of the VGGNET-16 shown in Figures 4 (b) and the configuration of VGGNET-16 is shown in figure (2).

4. Result Analysis

The dataset covers most of all Telugu characters in the literature approximately 1600 characters as shown in figure (2). It is the largest dataset available for Telugu characters as of now. Moreover, each character image is multiplicities with 12 using Image Data Argumentation (shifting, rotation, and sheering effect).

The experimental analysis is evaluated using Google Colab. VGGNET-16 algorithm is applied to the HandWritten Telugu Character dataset. This dataset covers most of all Telugu characters in the literature approximately 1600 characters.

For quantitative analysis of the experimental results, the following performance metrics are considered, including accuracy (AC), Dice coefficient (DC). To do this we also use the variables True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Confusion Matrix: it is a particular table format that allows the performance evaluation of an algorithm. The overall accuracy is calculated using an Equation.

\[
AC = \frac{TP + TN}{TP + TN + FP + FN}
\]

Table 4. Accuracy and loss

| Model # | Accuracy | Loss |
|---------|----------|------|
| Model 1 | 92%      | 0.5  |
| Model 2 | 80.5%    | 0.6  |
| Model 3 | 90%      | 0.2  |
| Model 4 | 50%      | 1.5  |

(a) (b)
Graph 1. (a,b) represents Model 1, (c,d) represents Model 2, (e,f) represents Model 3, (g,h) represents Model 4

In the above graph 1.(a), (c),(e), and (g) represents training loss and validation loss of model 1,2,3,4 respectively, and graph 1.(b),(d),(f), and (h) represents training accuracy and validation accuracy of Model 1,2,3,4 respectively, where x-axis represents number of epochs and y-axis represents accuracy and loss percentages. Model 1 achieved 92 percent accuracy and 0.5 loss. Model 2 achieved 80 percent of accuracy and 0.6 loss. 90 percent accuracy and 0.2 loss achieved by Model 3 where as Model 4 achieved 50 percent accuracy and loss of about 1.5.

4. Conclusion
In this paper, we are implemented in our own dataset and used a model called VGG-16. By using the VGG-16 we are divided the data into testing and training set. We trained and tested the data using VGGNET-16 and achieved good accuracy. By using this method we can directly recognize the Characters with no need to do segmentation and all because we scanned each and every character and created our own dataset and by using that dataset only we trained the network and we tested some documents and we can easily digitize the Telugu documents. This will help everyone who wants to store their Telugu documents into digital documents.
Acknowledgment:
The work is carried out Under the Collaborative Research project of TEQIP-III/JNTUH file no: JNTUH/TEQIP-III/CRS/2019/CSE/12 and Titled “Deep Learning Aided-OCR for Handwritten Telugu Character”.

References

[1] Atul Negi, Chakravarthi Bhagavathi, B. Krishna “An OCR System for Telugu”. IEEE, 2001, 0-7695-1263.
[2] C. Vasantha Lakshmi, C Patvardhan “A high accuracy OCR System for Printed Telugu Text”, IEEE,2003, 0-7803-7651.
[3] Atul Negi, Nikhil Kasinadhuni, Localization and Extraction of Text in Telugu Document Images”, IEEE,2003, 0-7803-7651.
[4] 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.
[5] Neerugatti Varipally Vishwanath, Manjunathachari and Satyaprasad, “Handwritten Telugu Composite Character Recognition Using Morphological Analysis”, International Journal of Pure and Applied Mathematics, 2018,119,18 2018, 667-676.
[6] Chakravarthy Bhagvati, Tanuku Ravi, S. Mahesh Kumar, Atul Negi,” On Developing High Accuracy OCR Systems for Telugu and Other Indian Scripts”, Proceedings of the Language Engineering Conference, IEEE,2002,0-7695-1885-0/02.
[7] Atul Negi, Nikhil Kasinadhuni,” Localization and Extraction of Text in Telugu Document Images”.
[8] Panyam Narahari Sastry, T.R.Vijaya Lakshmi, N.V.Koteswara Rao, T.V.Rajinikanth, Abdul Wahab, Telugu Handwritten Character Recognition using Zoning Features”, IEEE,2014, 978-1-4799-6541-0/14.
[9] Konkimalla Chandra Prakash, Y. M. Srikar, Gayam Trishal, Suraj Mandal, Sumohana S. Channapayya, “optical Character Recognition(OCR) for Telugu: Database, Algorithm, and Application”, IEEE, 2018, 978-1-4799-7061-2/18.
[10] K. Mohana Lakshmi, T Ranga Babu,” A New Hybrid Algorithm for Telugu Word Retrieval and Recognition”. International Journal of Intelligent Engineering and Systems, Vol.11, No.4, 2018.
[11] K. Vijay Kumar, R. Rajeshwara Rao, “Online Handwritten Character Recognition for Telugu Language Using Support Vector Machines”, International Journal of Engineering and Advanced Technology, ISSN: 2249 – 8958, Volume-3, Issue-2, December 2013.
[12] Rakesh Achanta, Trevor Hastie, “Telugu OCR Framework Using Deep Learning”,
[13] C.V.Chakradhar, B.Rajesh, M. Raghavendra Reddy, “A Study on online Handwritten Telugu Character Recognition”, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 5, Issue 7, July 2016.
[14] N. Shobha Rani, T. Vasudev,” Post-Processing Methodology for Word Level Telugu Character Recognition Systems using Unicode Approximation Models”,IEEE, 2015, 978-1-4673-6667-0/15.
[15] C. Vasantha Lakshmi, Ritu Jain, C. Patvardhan, “OCR of Printed Telugu Text with High Recognition Accuracies”,Springer,2006,786-795.
[16] Manpreet Kaur, Ashok Bathla, “A Review on Different Techniques for Character Segmentation for Hindi Handwritten Words”, International Journal Peer Reviewed Journal Refereed Journal Indexed Journal, 2454-6615.
[17] K.Karthick, K.B.Ravindrakumar, RFrancis, S.Ilankannan, “Steps Involved in Text Recognition and Recent Research in OCR; A Study” International Journal of Recent Technology and Engineering, 2277-3878, Volume-8, Issue-1, May 2019.
[18] Amit Choudhary, “A Review of Various Character Segmentation Techniques for Cursive Handwritten Words Recognition”, International Journal of Information & Computation Technology. 0974-2239 Volume 4, Number 6, 2014.
[19] Amjad Rehman, “Cursive Overlapped Character Segmentation: An Enhanced Approach”,
[20] Rajiv Kumar and Amardeep Singh, “Character Segmentation in Gurumukhi Handwritten Text using Hybrid Approach”, International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 201.
[21] A.Venkata Srinivasa Rao, Madasu Subbarao, Nenkkanti Venkata Rao, A.S.C.S.Sastry, L.Pratap Reddy, “Segmentation of Touching Handwritten Numerals and Alphabets”, 2009 Second International Conference on Computer and Electrical Engineering.
[22] S.E. Benita Galaxy, S. Selvin Ebenezer, “Enhancement of segmentation and zoning to improve the accuracy of handwritten character recognition”, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.
[23] Aravinda C.V, Dr.Prakash H.N, “A Review On Various HandWritten Cursive Characters Segmentation Techniques”, IEEE, 2016, 978-1-5090-2399-8/16.
[24] Li Chen, Song Wang, Wei Fan, Jun Sun, Satoshi Naoi, “Cascading Training for Relaxation CNN on Handwritten Character Recognition”, 2016 15th International Conference on Frontiers in Handwriting Recognition.
[25] Akm Ashiquzzaman and Abdul Kawsar Tushar, “Handwritten Arabic Numeral Recognition using Deep Learning Neural Networks”, IEEE, 2017, 978-1-5090-604-7/17.
[26] Kian Peymani, Mohsen Soryani, “From Machine Generated To Handwritten Character Recognition; A Deep Learning Approach”, 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA 2017) April 19-20, 2017.
[27] Martin Rajnoha, Radim Burget, Malay Kishore Dutta, “Handwriting Comenia Script Recognition with Convolutional Neural Network”, IEEE,2017, 978-1-5090-3982-1/17.
[28] Ruyu Zhang, Qingqing Wang, Yue Lu, “Combination of ResNet and Center Loss Based Metric Learning for Handwritten Chinese Character Recognition”, 2017 14th IAPR International Conference on Document Analysis and Recognition.
[29] Bishwajit Purkaystha, Tapos Datta, Md Saiful Islam,” Bengali Handwritten Character Recognition Using  Deep Convolutional Neural Network”, International Conference of Computer and Information Technology (ICCIT), 22-24 December 2017.
[30] Mohammad Agung Wibowo, Muhamad Soleh, Winangsari Pradani, Achmad Nizar, “Handwritten Javanese Character Recognition using Discriminative Deep Learning Technique”, International Conferences on Information Technology, Information Systems and Electrical Engineering.