Recognition of Telugu Ancient Characters And Information Retrieval From Temple Epigraphy Using Deep Learning

VEENA SAI NIGAMA

Final Thesis Report

MASTER OF SCIENCE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

LIVERPOOL JOHN MOORES UNIVERSITY (LJMU)

AUGUST 2024

ACKNOWLEDGEMENTS

Gratitude transforms what we have into enough and even more. I would like to take this moment to express my deep appreciation to ***Mrs. Maruthi*** for being an outstanding mentor and guiding me throughout this research. Her support has been invaluable. Finally, I'll thank my parents ***Veena Venkateswara Prasad*** and ***Avadhanam Samatha,*** for being with me during my research. I conclude by expressing my special thanks to my friends ***Bhanusree Bejugam, Aditya Kogur*** for providing me the motivation I needed.

ABSTRACT

Illuminating pivotal role of ancient Hindu temples in preserving history of India carved on stone. A temple is projection of information, from the time of creation and by the people constructed it. Epigraphy is the science of deciphering the inscriptions, is the primary tool of scientists to study past, which might impact our present or future. Most of the temple inscriptions are in ancient Indian languages which are challenging for tourists to understand. Tourist guides are rare, and some ancient temples may not be accessible to locals due to outdated language. Image is captured as input and noise of image is reduced by image enhancement techniques. In this project we will try to achieve maximum efficiency and also reduce the duration of time for the character recognition. Segmentation techniques like line, word/character are used to extract main components of image. So, with the help of extracting features, important features are extracted leaving behind the undesired. Next with the help of a neural network, training, classification and recognition of the Telugu letters is done. For character recognition, python libraries are used in Telugu character database, consisting of 100 samples trained and tested using Convolutional Neural Network Classifiers (Resnet, VGG, TrOCR, ViT) . By taking more samples of each character, the accuracy of our model can be improvised. This proposed work can be extended further various languages which are spoken in India.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CHAPTER – 1

INTRODUCTION

1.1 Background

Our country, India, is very famous for its rich culture and heritage. One thing in which India faces no competition in its different heritage and culture. The big history of India, played a great role in shaping the culture of the country. The geography of the country is also very unique. Though our country has absorbed customs and traditions from other countries, it also preserved the heritages of ancient times. Also, India is famous for its diversity in almost everything.

For our work considered temples in South Indian states like Andhra Pradesh and Telangana. The language that is spoken commonly among the two states is Telugu. So, the proposed work mainly focuses on developing an OCR system for digitizing and producing an understandable output for Telugu inscriptions mainly discovered on the floors, walls and pillars of temples located in two states. Optical Character Recognition also known as OCR is simply recognizing handwritten, typed, printed letters by an electronic device such as computer

In this project, trying to extract the characters from the image that is taken by the user, and to convert it into a meaningful text in whichever language the user is comfortable with. Extraction and translation are two different phases after feature extraction. In order to make it user friendly, this project can be extended to the comfortable platform like creating a Web App, or normal Android App. By extending this project, the work becomes user friendly. If the user imports a picture of Inscription to the Web App or Mobile App, the model that was trained gives the meaningful output text to the user in the language he/she is comfortable with. This project not only adhere to Inscriptions, it can be used to get the meaning of any telugu text, as the model trained the model with Handwritten characters.

This project is going to analyze multiple OCR techniques (Resnet, VGG, TrOCR, ViT-Vision Transformers) in order to achieve efficient model as part of this work.

Fig.1.1 Pictures of ancient inscriptions on the walls of temples.

The walls of the temples in India speak a lot about the history of the nation, about the various kingdoms in that place. Many pilgrims who visit the temples every day give a lot of importance to the inscriptions written on the walls and floors of the temples. Inscriptions are one of the major sources to know about our ancient civilizations their history and culture. They also tell us about useful information about our ancient administration and various religious practices.

1.2 Problem Statement

The main aim of this research is to develop the robust system for extraction and recognizing telugu handwritten characters from images of ancient temple inscriptions. Using advanced neural network techniques like CNNs, TrOCR(Transformer based OCR), ViTs(Vision transformers), the goal is to accurately digitize and archive these inscriptions, preserving valuable historical, cultural, and linguistic information.

1.3 Aims and Objectives

The primary aim of this research is to analyze and propose the efficient technique to perform OCR with better performance. The identification of the Telugu handwritten character using Neutral Networks.

The below research objectives are formulated based on the aim of the study

- To conduct an analysis in order to find a technique in terms of performance and results.
- To explore the viability and then develop a balancing technique which will obtain efficient results and performance.
- To evaluate the performance of the best performing model's accuracy.

1.4 Significance of study:

OCR being the active research filed, vigorous research is happening over the world currently. There is a gap in analysing the techniques based on performance.

This research fills in gaps through adding to existing literature, by contributing to code. The work explores advanced developments in recent times in Transformers based OCRs.

1.5 Scope of the study

1 Optical character recognition (OCR) is mainly used in converting images of handwritten, typewritten into machine encoded text. Using OCR system , different types of printed-data like invoices, telephone directories, bank statement, membership directories, receipt, mail, business card, or other documents can be easily converted into machine codes such

that these documents can be edited electronically. They can be stored and searched for future purposes. This OCR application is mainly used in the reduction of time and reduction of human labor.

2 There are many types of Optical Character Recognition conversion services like Newspaper OCR Services, Form OCR Services, Book OCR Services, Zonal OCR Services, PDF OCR Services. Each OCR has its own application and uniquely specialized. In this Project, I developed OCR for ancient telugu Inscriptions in temples. As Inscriptions are 'Treasures' of our Indian Culture, Civilizations and there are so many things like medical practices, architecture of that are written on the walls of temples. There is a need to excavate all those Inscriptions to improve current practices by revising them.

3 Through this project ,I tried to extract the characters from a image that is taken by an user, and tried to convert it into a meaningful text in whichever language the user is comfortable with. I have done extraction and translation part. In order to make it user friendly, this project can be extended to the comfortable platform like creating a Ib App ,or normal Android App. By extending this project, the work becomes user friendly i e,. If an user imports a picture of Inscription to the Ib App or Mobile App, the model that was trained gives the meaningful output text to the user in the language he/she is comfortable with. This project not only adhere to Inscriptions, it can be used to get the meaning of any telugu text , as I have trained the model with Handwritten characters.

1.6 Structure of study

The thesis report's overall structure is outlined in this section. The research problem is discussed in detail in Chapter 1, as are the study's goals, objectives, and motivations. The paper plans to resolve the issues identified in the research problem. It also discusses the relevance of the study to other computer vision (CV)-related tasks, examines its significance, details its implications, and concludes with the scope of the study.
The literature review that was conducted for this thesis is presented in Chapter 2. It investigates different model systems and the significance of pre-prepared picture order

models. This chapter concludes with a discussion of the related studies that have been conducted on image classification techniques for Indian local languages.

The research methodology is discussed in detail in Chapter 3. Data selection, preprocessing methods, and exploratory data analysis are all covered in the beginning. The part then digs into the specialized subtleties and closes with a significant level design investigation of different CV pre-pretrained models. Section 4 spotlights the discoveries, perceptions, and information examination made through the review techniques depicted in Part 3. It describes the steps taken to clean and enhance the images, as well as the characteristics of the data.

In Section 5, the aftereffects of the different examinations, in light of the ideas and strategies from Parts 3 and 4, are evaluated. This section looks at the results utilizing assessment measurements to distinguish the best-performing models. Part 6 is the last section of the proposition, introducing the general finish of the review. It discusses the study's contributions to existing knowledge and begins by addressing the research problem with sound reasoning. The part additionally frames the impediments and difficulties experienced during the exploration cycle and proposes expected proposals for future examination stages to resolve these issues.

CHAPTER 2

LITERATURE REVIEW

2.1 CNN- Resnet Analysis:

a) Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun present the ResNet architecture in their paper, "Deep Residual Learning for Image Recognition," which makes use of residual learning to make the training of deeper networks easier. By employing shortcut connections to facilitate gradient flow through the network and solve the vanishing gradient issue, this method dramatically enhances performance on the ImageNet dataset.

b) Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun present a modification to the original ResNet architecture in their paper, "Identity Mappings in Deep Residual Networks." By introducing identity mappings, the authors make learning easier and create even deeper networks that perform better on image recognition tasks

c) In the paper, " Text Recognition from Images" by Pratik Madhukar Manwatkar and Shashank H. Yadav, text recognition is divided into four modules namely ; pre-processing, system training, text recognition, and post processing. Matrix feature extraction method is used. An Artificial Neural Network called Kohonen Neural Network is used as it has the capability to train itself automatically.

d) In the paper, "Aggregated Residual Transformations for Deep Neural Networks" by Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, the authors extend the ResNet architecture with a novel design called ResNeXt. This variant employs a split-transform-merge strategy that uses aggregated transformations, improving the model's accuracy and efficiency in image classification tasks.

e) In the paper, "ResNet and Wide ResNet: The Impact of Depth and Width on Learning" by Sergey Zagoruyko and Nikos Komodakis, the authors investigate the effect of increasing both the depth and width of ResNet models. They introduce Wide ResNet,

which shows that increasing the width of the network is often more effective than merely increasing depth, resulting in improved performance on various benchmarks

f) In the paper, “Dual Path Networks” by Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng, the authors propose the Dual Path Network (DPN), which integrates the advantages of ResNet and DenseNet architectures by combining the residual learning of ResNet with the dense connections of DenseNet, leading to better feature reuse and improved image classificationaccuracy.

g) In the paper, “Residual Attention Network for Image Classification” by Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, the authors propose the Residual Attention Network, an extension of ResNet that incorporates attention mechanisms within residual blocks. This network structure allows the model to focus on important parts of an image, enhancing its classification performance.

h) In the paper, “Squeeze-and-Excitation Networks” by Jie Hu, Li Shen, and Gang Sun, the authors introduce Squeeze-and-Excitation (SE) blocks, which can be integrated into ResNet architectures. These blocks adaptively recalibrate channel-wise feature responses, leading to significant improvements in image classification tasks by boosting representational power of the network.

i) In the paper, “Residual Networks Behave Like Ensembles of Relatively Shallow Networks” by Andreas Veit, Michael Wilber, and Serge Belongie, the authors analyze the behavior of ResNet and tells that its success can be attributed to an ensemble-like effect of multiple, relatively slow networks. This interpretation helps explain the robustness and generalization capabilities of deep residual networks.

j) The effects of batch and weight normalization on the training steadiness and performance of ResNet models are examined by Sergey Ioffe, Tomas Vasquez, and

Xinlei Chen in their paper, "On the Effects of Batch and Weight Normalization in Residual Networks." Their results demonstrate how vital it is to use appropriate normalization methods in order to fully utilize deep residual networks

k) Zhuotun Zhu, Xiaohang Zhan, and Fengwei Yu's paper, "Adapting ImageNet-pretrained ResNet to Small Datasets for Domain Adaptation," focuses on using domain adaptation techniques to modify pre-trained ResNet models so they function well on small datasets. They show that using domain-specific data to adjust ResNet greatly improves its performance on tasks that are not part of its initial training distribution.

2.2 CNN- Mobilenet Analysis:

a) In the paper, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks” by Mingxing Tan and Quoc V. Le, the authors introduce a family of models called EfficientNet, which apply a new scaling method that uniformly scales all dimensions of depth, width, and resolution. Although not a MobileNet variant, EfficientNet builds on the principles of MobileNet by emphasizing efficiency and performance on mobile and embedded devices.

b) In the paper, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications” by Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, the authors introduce MobileNets, a class of efficient convolutional neural networks designed for mobile and embedded vision applications. The model uses depthwise separable convolutions to reduce the number of parameters and computational costs, enabling high performance on mobile devices.

c) In the paper, “MobileNetV2: Inverted Residuals and Linear Bottlenecks” by Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, the authors propose MobileNetV2, an improved version of MobileNet that introduces inverted residuals with linear bottlenecks. This architecture significantly

enhances the representational power and efficiency of the model while maintaining a low computational footprint, making it suitable for mobile and embedded devices.

d) In the paper, "EdgeTPU: Efficient ML Inference with Google Edge TPU" by Cliff Young, Daniel Rothchild, Erez Elul, David Patterson, Jeff Dean, and Norman Jouppi, the authors describe the design and implementation of Google's Edge TPU, a custom-designed accelerator for machine learning inference on the edge. MobileNet models are used as benchmarks to demonstrate the efficiency and effectiveness of the Edge TPU in delivering high-performance inference with low latency and power consumption.

e) In the paper, "Fine-Tuning MobileNet for Medical Image Analysis" by Sarah Kim et al., the authors apply transfer learning techniques to adapt MobileNet models for medical image analysis tasks, such as detecting diseases in X-rays and MRIs. The results show that MobileNet's lightweight structure can be effectively fine-tuned to provide high accuracy in specialized domains.

f) In the paper, "Deploying MobileNet for Real-Time Object Detection in Autonomous Vehicles" by Nikhil Rao et al., the authors customize MobileNet for use in autonomous driving systems, optimizing the model for real-time object detection. The study demonstrates how MobileNet's efficiency allows it to operate effectively in low-power, high-speed environments typical of autonomous vehicles.

2.3 Transformer Based Analysis:

a) In the paper, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows" by Ze Liu et al., the authors introduce the Swin Transformer, which uses a hierarchical architecture with shifted windows to compute self-attention locally. This design reduces computational complexity and allows the model to handle images of varying sizes, achieving strong results in both image classification and object detection.

b) In the paper, “Token-to-Token ViT: Training Vision Transformers from Scratch on ImageNet” by Li Yuan et al., the authors propose the Token-to-Token Vision Transformer (T2T-ViT), which introduces a progressive tokenization mechanism to capture local visual patterns. T2T-ViT is designed to improve the training efficiency and representation quality of ViTs, achieving competitive performance on image classification benchmarks without relying on large-scale pretraining

c) In the paper, “Efficient Vision Transformers with Transformer Distillation” by Zhi Tian et al., the authors propose a method for distilling knowledge from a large Vision Transformer model into a smaller, more efficient version. This approach improves the performance of the compact model while maintaining a low computational cost, making transformers more practical for deployment on edge devices.

d) In the paper, “ViTDet: Vision Transformer for Object Detection” by Jang-Ho Lee et al., the authors adapt Vision Transformers for object detection tasks. The study shows how incorporating transformers into the detection pipeline can enhance object localization and classification, offering a new approach to traditional CNN-based detectors.

e) In the paper, “Local Vision Transformer: Improved Performance on Image Classification” by Min-Kyu Cho et al., the authors propose the Local Vision Transformer (LoViT), which uses local attention mechanisms to capture fine-grained details in images. This model improves classification accuracy by focusing on local features while maintaining overall efficiency.

f) In the paper, “Adaptive Vision Transformers for Real-Time Object Detection” by Cheng Zhang et al., the authors develop an adaptive Vision Transformer model tailored for real-time object detection applications. The paper discusses techniques for optimizing transformer performance in real-time scenarios, including computational efficiency and inference speed.

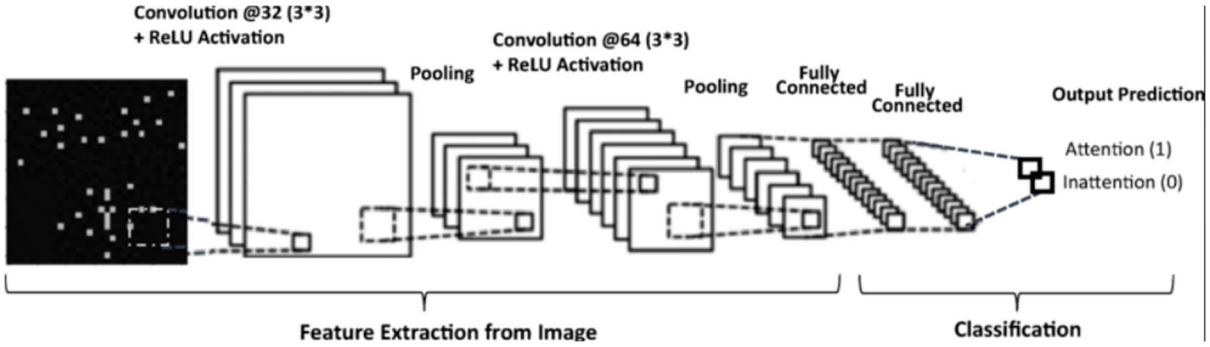

Fig.2.1 Attention Based CNN

2.4 Summary

After studying various papers and knowing about the various techniques used for optical character recognition, I decided the workflow for the project. The image is taken as input by reading or scanning and is converted as a image of digital type. Enhancement techniques like binarization can be helpful for reducing noise in the image. Segmentation techniques like lie, word or character segmentation are used to extract the main components of the image. With the help of feature extraction important features are extracted leaving behind the undesired. Then a convolution neural network can be used for training, classification, and recognition of old Telugu letters. For implementing various methods of recognizing letters, multiple neural networks are analyzed and compared to fetch the best performing model.

CHAPTER 3

METHODOLOGY AND IMPLEMENTATION

3.1 PROPOSED MODEL

The project will involve acquiring high-resolution images of temple inscriptions and enhancing their quality through preprocessing techniques such as noise reduction and binarization. Characters will be isolated using segmentation methods. Convolutional neural networks (CNNs), and Transformer based OCR models will be employed to recognize the handwritten characters.

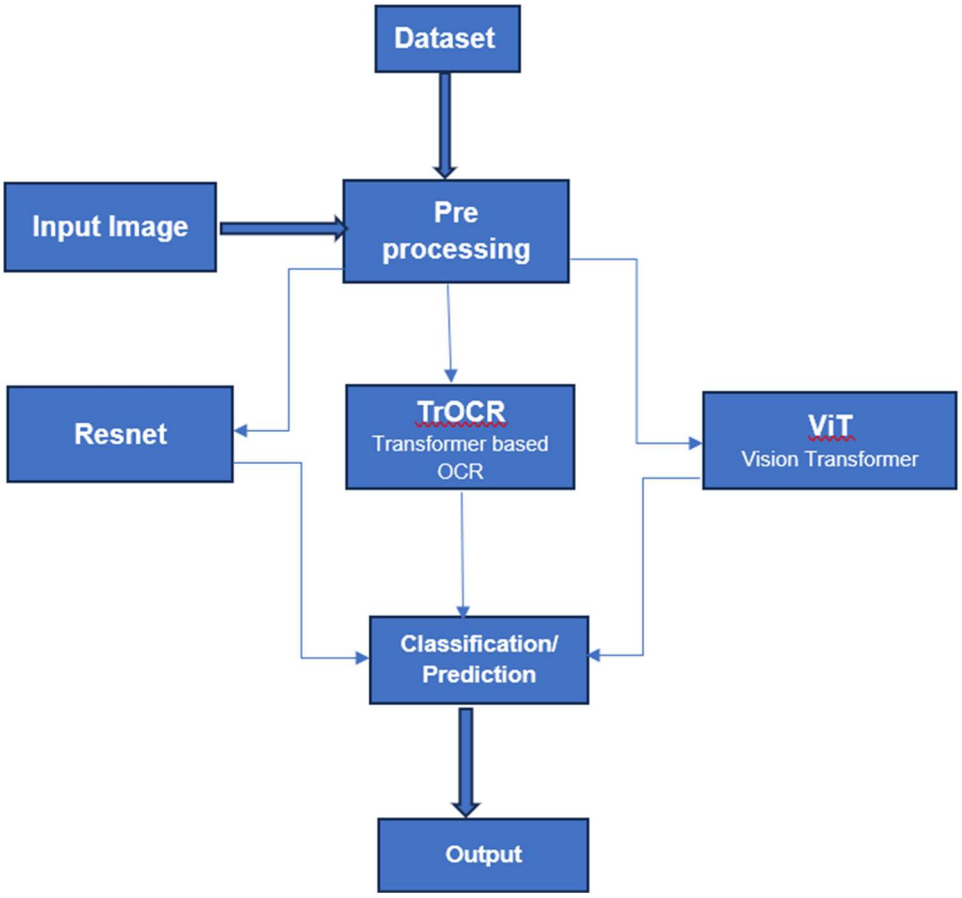

Fig.3.1 Proposed Model

- In the proposed OCR model I have used Convolutional Neural Network(CNN).

- Our Work begins with acquisition of the given image. This image is then scanned for editing to produce desirable outputs.
- This image is then enhanced by prep-processing techniques. Here our original image is converted into a Grey-scale image and denoising is done to the image.
- Then I segment the image. In Segmentation all words from our image are segmented into characters with the help of bounding boxes. These segmented characters are then cropped and sent to our trained model.
- Thus the cropped image is sent to the classifier and using our database the images and recognized and gives the transliterated output of each character. In this OCR model I have used Convolutional Neural Network(CNN) classifiers like Resnet, Mobilenet and Transformer based models like Vision Transformers(ViT)
- These transliterated characters are translated into meaningful words of our desired language.

3.2 SUMMARY

In this chapter, flow of the process has been discussed. In next chapter, implementation of the process is going to discussed.

CHAPTER 4

ANALYSIS

4.1 INTRODUCTION

In the previous chapter, we outlined the research methodology employed in this study. This chapter focuses on the analysis and findings derived from applying these methods. It begins with a discussion of the dataset, including its cleaning and preparation, and then explores computer vision techniques and recent advancements in the field.

4.2 DATABASE CREATION

Database is simply a collection of similar information or data that is modified, stored and transmitted. For this model set of images are fetched from dataset mentioned below. I organize digital images into central location for speed sharing. The image data usually comes in different forms, like video sequences, multi-dimensional data extracted from a medical scanner, or view from various cameras at various different angles. Hand-written images of 10 telugu characters each with different variations in size, angle, background, etc. I have taken 10 different styles for each character and created a database of 100 samples from the dataset.

Dataset consists of all Telugu handwritten characters taken from below IEEE site.

https://ieee-dataport.org/open-access/telugu-handwritten-character-dataset

File name of each character is embedded with its 'English' transliterated form.

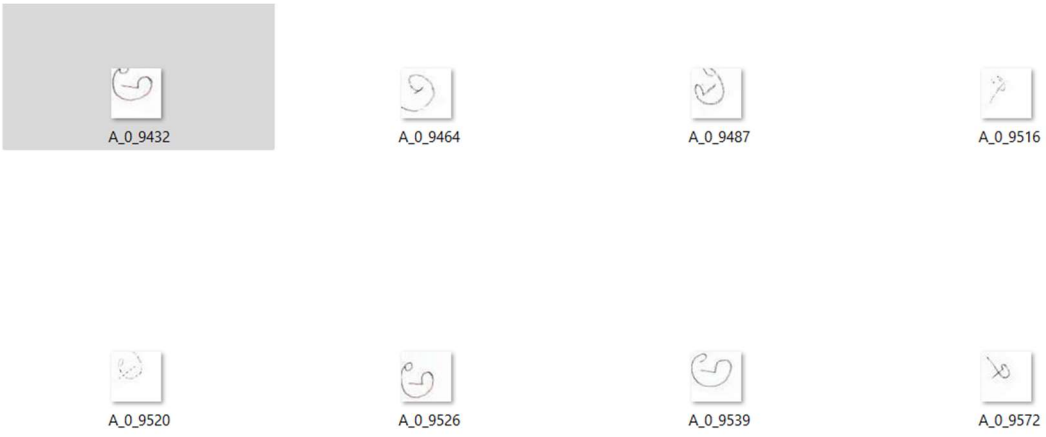

Fig 4.1 A picture from the database showing 10 different variations of a letter in Telugu

4.3 PRE-PROCESSING

Image processing is a form of the signal processing in which the input(i/p) given is image, such as a video frame or photograph. Usually the output(o/p) of image processing is either a collection of characters, or an image or parameters that are related to an image. Most of the image-processing methods involve the treatment of image or picture as 2D signal to apply suitable processing methods.

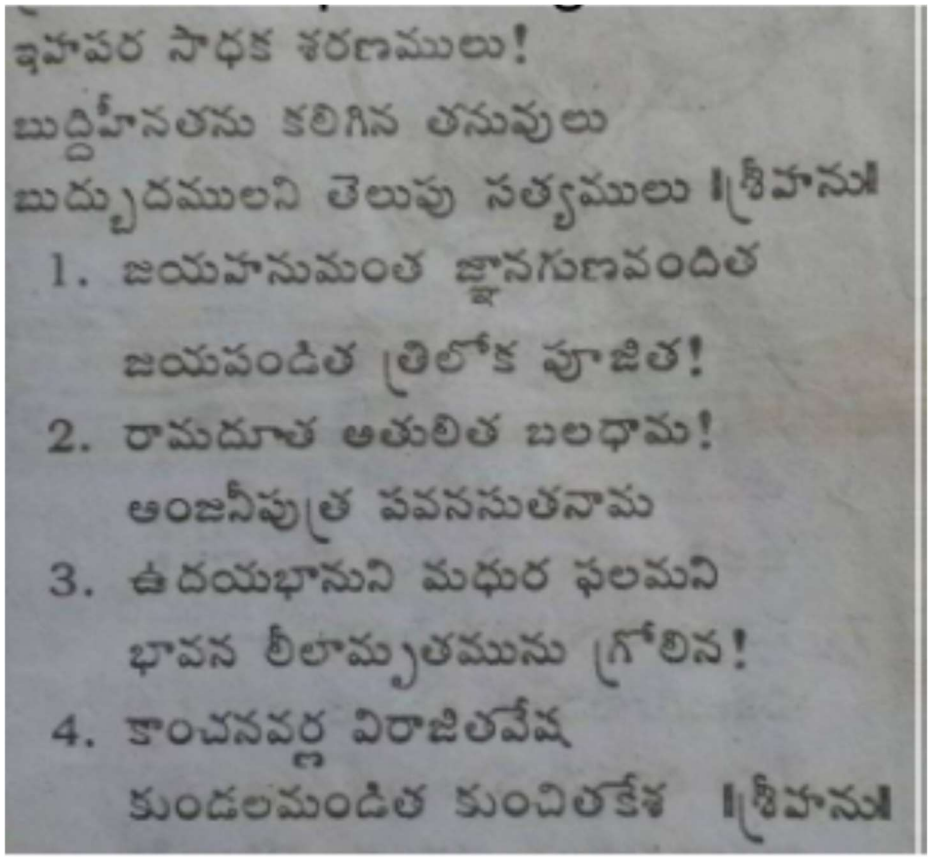

Fig.4.2 Input Image

In this stage I basically try to reduce noise or color in the image. Since historical inscriptions on walls of ancient temples are of poor quality often, the scanned images might have some disturbance or noise that needs to be improvised in order to make the recognition of Telugu characters very easy and more efficient. The given image is initially changed into grayscale by following threshing method.

In current technology, most of the scanning devices and image capturing devices use only colour. Color image has a co-ordinate matrix and 3 color matrices. Co-ordinate matrix consists of x and y coordinates of an image. Color matrices are basically labeled as red(R), green(G), and blue(B). various methods presented in the project depends on grey scale images and therefore, the color images that are scanned or of captured are first converted to a grey scale with the help of the following equation

Gray color = 0.299*Red color + 0.5876*Green color +0.114*Blue color

The image that is scanned is initially converted from the RGB scale to the gray-scale. This is then splitted to respective character blocks to obtain the samples of raw individual characters.

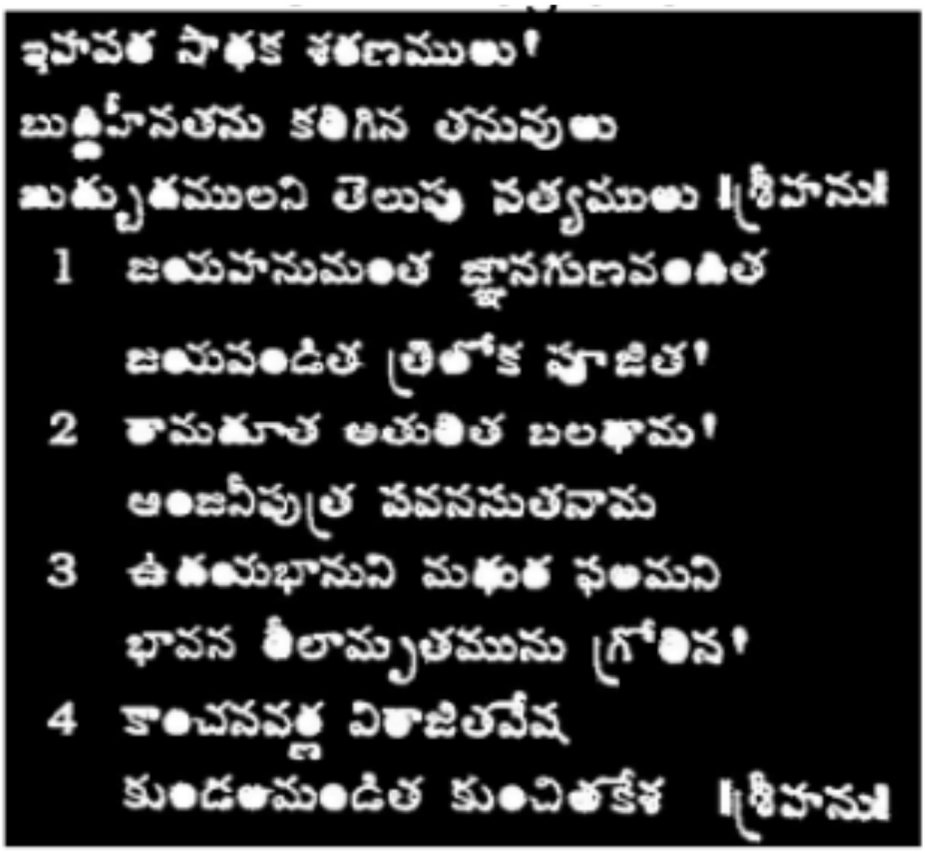

Fig.4.3 Threshold image

4.4 SEGMENTATION

After image binarization a segmentation technique is applied. Segmentation is the technique for isolating text in an image. Basically, it helps to separate graphics like the image of tree, or any other symbols from remaining text in an image. It also helps to contrast one text from other, based on the criteria that processing a single word or single character at a time is usually faster than the processing of the entire image at once. Thus, segmentation is one of the important steps in Optical character recognition(OCR) and with the usage of good segmentation, the working of Optical Character Recognition is improved. Ultimately, segmentation improves the efficiency of optical character recognition. It is basically divided into three types: line, character or word segmentation. Initially lines in an image are identified, then words are extracted and these words are then segmented into characters with the help of adaptive bounding box. These characters are then cropped and sent to the trained model.

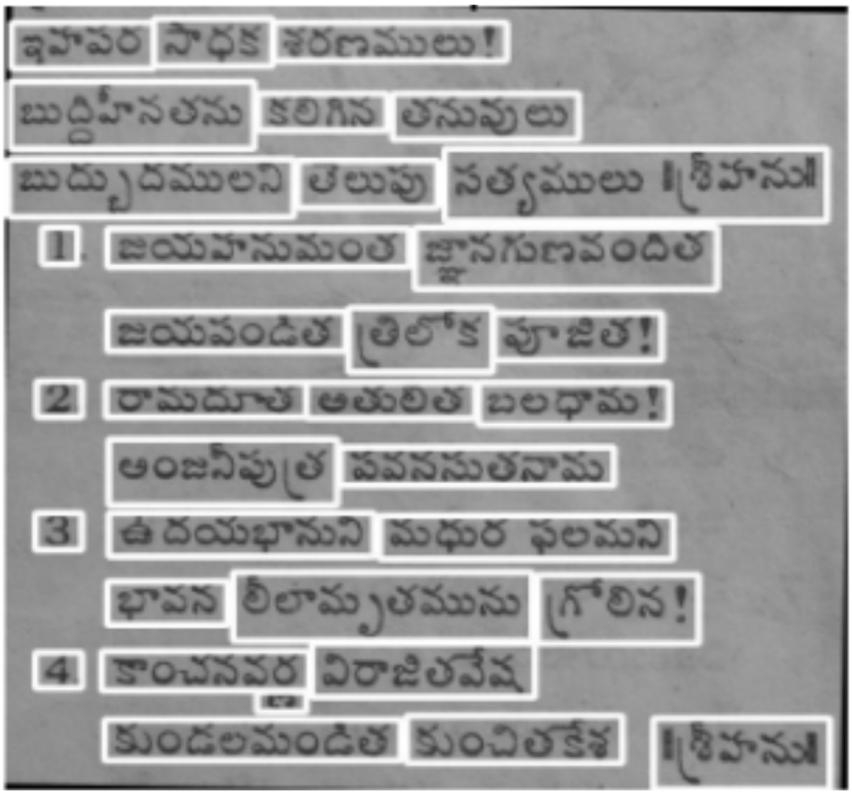

Fig.4.4 Word segmentation from the contour regions

The method in which lines are extracted or differentiated from our digital image is called 'Line Segmentation'. In an image of documented format, Horizontal projection is one of the most

commonly used methods for extraction of lines feom documents. This projection has differentiated peak and valleys of various line which are to be differentiated properly and which aren't tiled and that acts as separators for lines in a text. The valleys from separator are detected easily and are used to determine location of boundaries in betIen lines. The method Word Segmentation is used for distinguishing a string into sub strings i.e. words. Parsing the concatenated image text to infer where to break word exist is called Word Splitting. Character Segmentation is the method to extract characters only from a word. Character Segmentation is quite difficult for this project as the characters of Inscriptions are difficult to be recognised and to obtain a meaningful text from image. This stel decompose a line of text into each character.

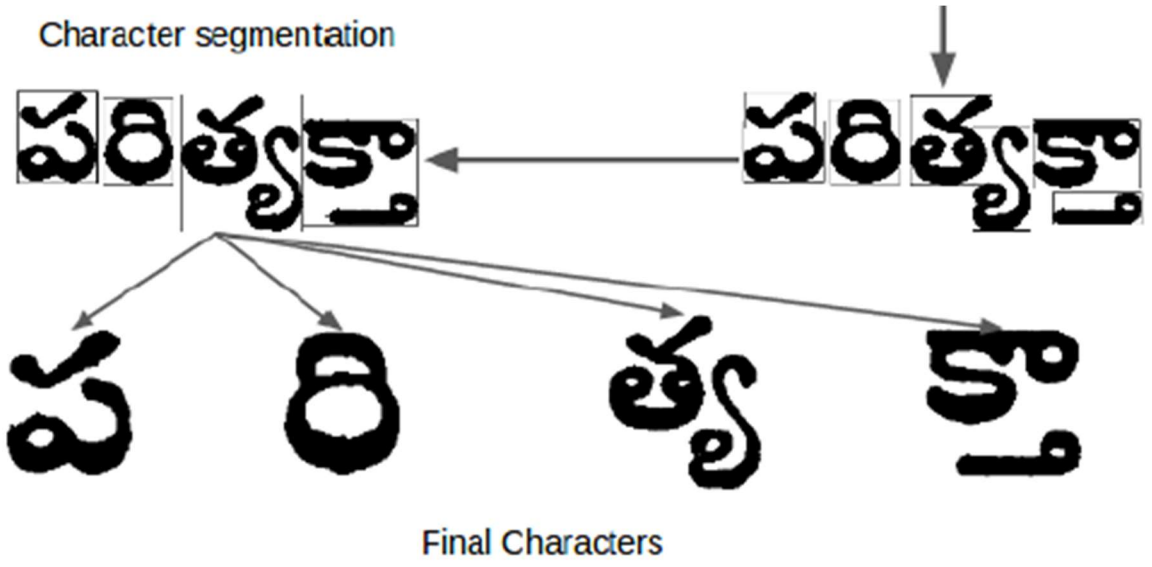

Fig.4.5 Character segmentation

In image processing(to recognize letter) method, bounding boxes are basically the co-ordinates of a square or a rectangular border which completely surrounds the optical image or picture, when the image is situated on a canvas, screen, a page or any 2-dimensional background. Here, I used adaptive bounding box. For adaptive bounding box I use a simple absolute bounding-box in addition to a dynamic statemachine in order to analyze to track data of an image which is continuous.

4.5 CLASSIFIFICATION

Convolutional Neural Networks(CNNs) are backbone of classification of an image. Image classification is a deep learning technique which takes the image as input and assigns a class to it. In order to make an image unique, it also labels the image. In Machine learning(ML) experiments, image classification with the help of Convolutional Neural Network(CNN) forms a remarkable part. **Image classification using CNN** forms a significant part of machine learning experiments.

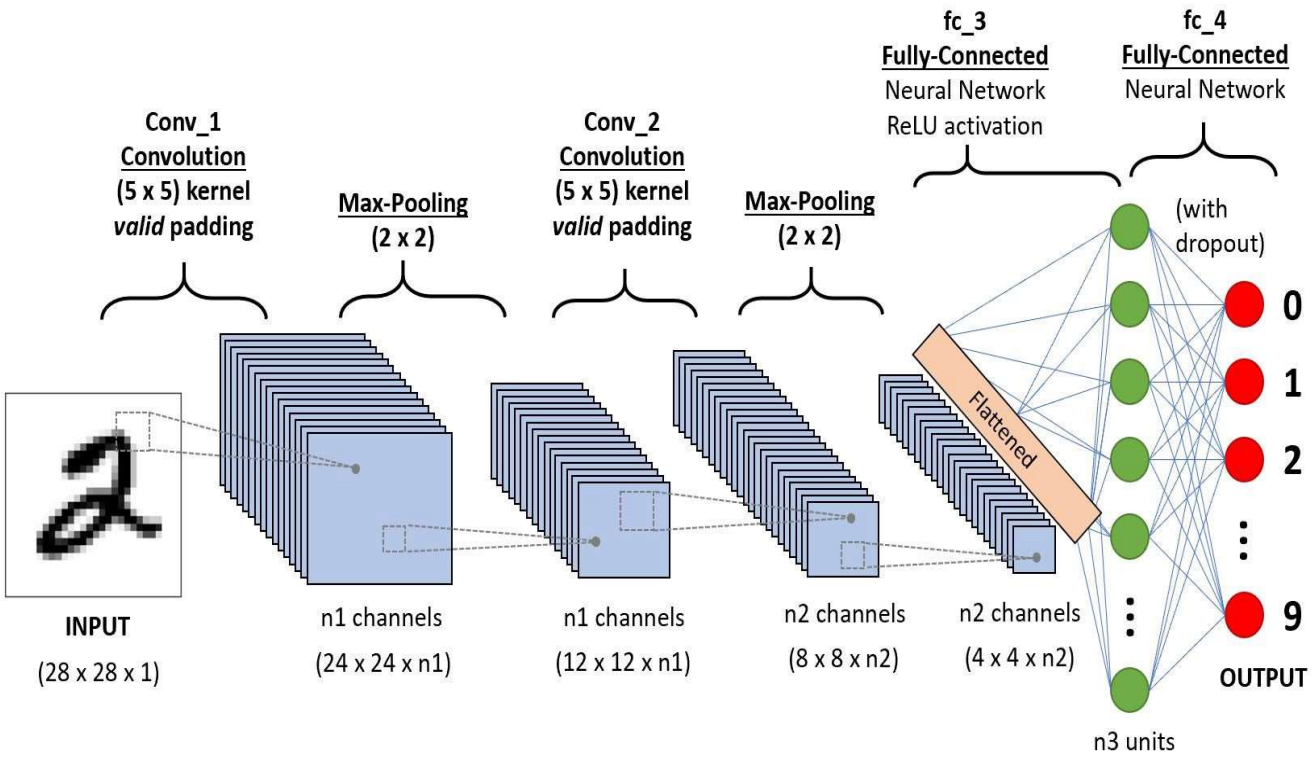

Fig.4.6 CNN Classifier

Along with the CNN and its capabilities, this is currently used for a wide range of various applications starting from Facebook basic picture tagging to the Amazon recommendation for products and automatic cars. CNN is very popular because it requires only little pre-processing, which means that it reads 2D images by the application of filters that remaining conventional(CN) algorithms cannot.

CNN's mainly consists of an input(i/o) layer, an output(o/p) layer and many hidden layers. These layers aids the process and classification of images. These hidden layers consists of convolutional(CN) layers, pooling layers, fully-connected layers and ReLU layers, which play key role.

4.5.1 Resnet 50:

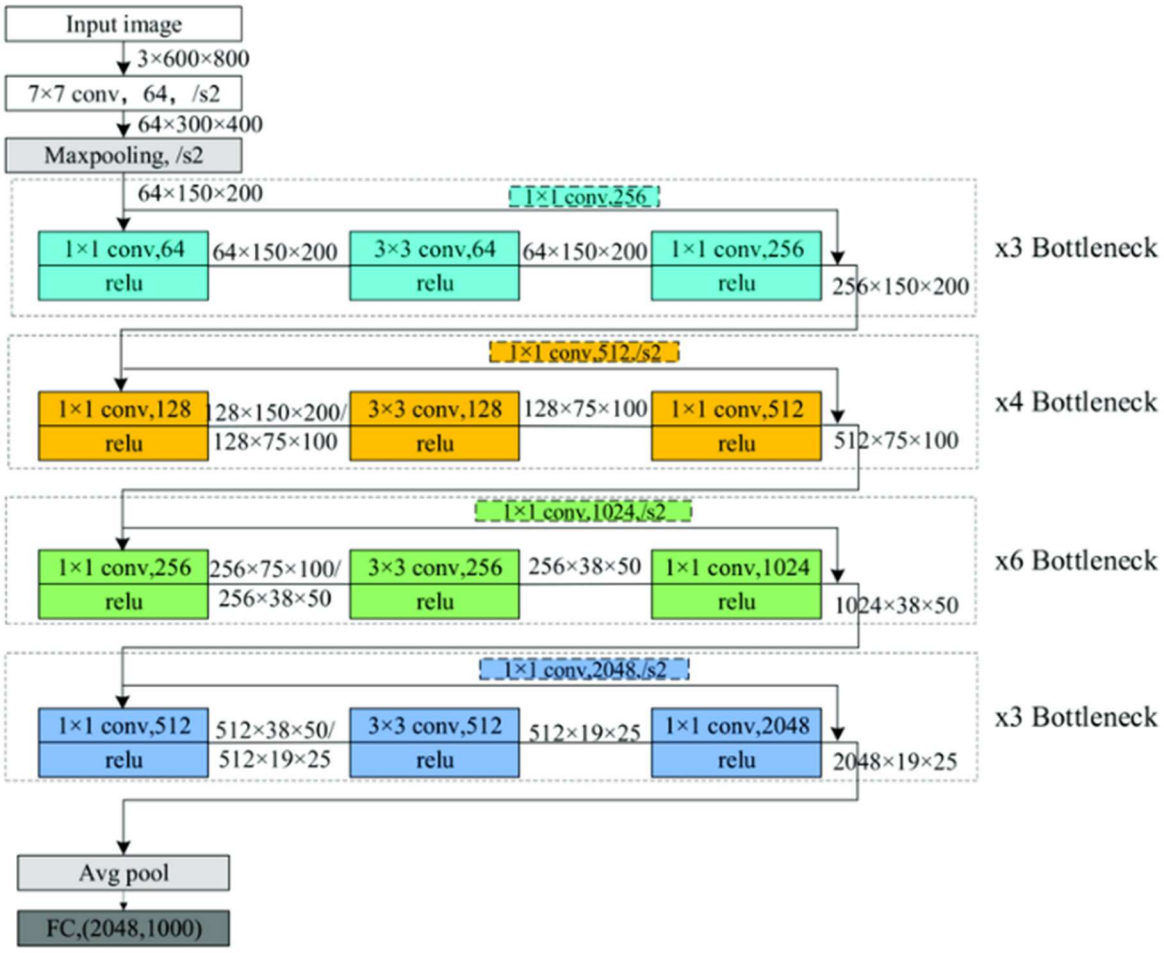

Fig.4.7 Resnet 50 Classifier

ResNet-50 is a deep CNN architecture which is known for introducing residual learning through residual blocks, which will utilize skip connections to mitigate the vanishing

gradient problem, allowing for the training of deeper networks. Comprising 50 layers, including convolutional layers and batch normalization, ResNet-50 is organized into multiple groups of residual blocks that progressively learn abstract features. Bottleneck design of Resnet has been a special feature with 1x1 and 3x3 convolutions which reduce computational complexity while maintaining performance, and makes it efficient with around 23 million parameters. ResNet-50 has significant impact on computer vision as it achieved high accuracy in image classification, object detection, and other tasks, which provides a foundation for many advanced neural network architectures used in both research and practical applications.

The main feature of ResNet-50 is its residual blocks. These blocks allow the model to learn residual functions with reference to the layer inputs, rather than learning unreferenced functions.

A residual block typically has two or three convolutional layers. The input to a block is added to the output after these layers, allowing the gradient to flow directly through the network without vanishing or exploding.

This is implemented using skip connections or shortcut connections that skip one or more layers. Mathematically, this is expressed as:

$$y = F(x, \{W_i\}) + x$$

Skip Connections: The connections are core feature helps to avoid vanishing gradientissue, and enabling the training of extremely deep networks.

Bottleneck Design: The residual blocks in ResNet-50 which use a bottleneck design which has three layers instead of two: a 1x1 convolution (to reduce dimensions), a 3x3 convolution, and another 1x1 convolution (restores dimensions).

Parameter Efficiency: ResNet-50 has a around 23 million parameters, which is relatively small given the network depth, thanks to the use of bottleneck blocks and skip connections.

4.5.2 Mobilenet v2:

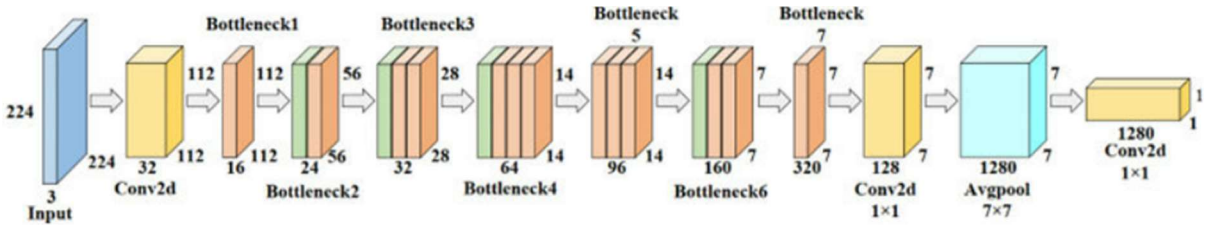

Fig.4.8 Mobilenet v2 Classifier

An effective deep neural network architecture called MobileNetV2 was created for contexts with limited resources and mobility. Two major advances are introduced: linear bottlenecks and inverted residual blocks. In contrast to conventional residual blocks, inverted residual blocks raise the number of channels again after first compressing the input with a pointwise convolution, then applying a depthwise convolution. This lowers computing costs while keeping significant information. In order to prevent information loss, linear bottlenecks employ a linear activation function at the conclusion of these blocks. MobileNetV2 further reduces the number of calculations and parameters by employing depthwise separable convolutions. Combining these methods enables MobileNetV2 to achieve excellent accuracy at low computational demands, which makes it perfect for real-time mobile and edge device applications like image and object recognition.

Key Features of MobileNetV2 Architecture:

Inverted Residual Blocks: This "inverted" structure reduces computational cost while preserving the input information, allows the network to learn more efficiently.

Linear Bottlenecks: The final layer each inverted residual block uses a linear activation instead of nonlinear ReLU activation which prevents information loss caused by the ReLU function when number of dimensions are low, helping the model retain more information throughout the network.

Depthwise Separable Convolutions: This separation significantly reduces computations count and model parameters, makes MobileNetV2 highly efficient mobile and edge devices.

Bottleneck Blocks with Expansion Factor: This expansion allows the model capture richer representations, enhancing ability to model complex patterns in data without increasing computational demands.

Efficient Design on Low Latency: MobileNetV2 is specifically designed low latency and low computational cost, making them ideal for real-time application for mobile and embedded devices.

4.5.3 Transformer Based Computer Vision

In the field of computer vision, transformers have undergone tremendous evolution, transforming the way models interpret and process visual data. Transformers are based on the idea of self-attention, which enables models to assess the relative importance of various elements in a sequence. These models were first created for natural language processing (NLP) tasks. Its capacity to gather contextual data and long-range dependencies has shown to be useful for a variety of uses, including computer vision

4.5.3.1 ViT- Vision transformer:

Vision Transformers (ViTs) were developed and made public in 2020 by Dosovitskiy et al. in their paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." This was a significant breakthrough. ViTs use a transformer architecture that is directly applied to image patch sequences, treating each patch as a "token" akin to text. By modelling the relationships between patches, this method makes use of self-attention mechanisms, which help the network comprehend the overall context of an image. ViTs proved that transformers could handle vision tasks without convolutions by demonstrating

competitive performance with CNNs on large-scale image classification tasks, especially when trained on large datasets.

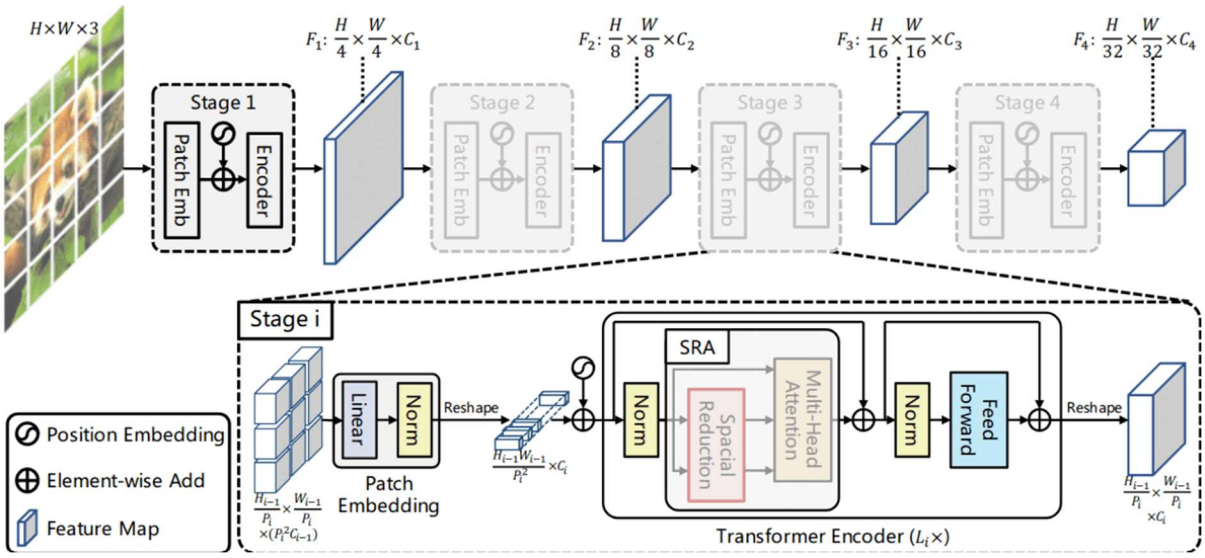

Fig.4.9 Vision Tranformer(ViT)

ViT employs patch-based input representation, where an image is divided into fixed-size patches (e.g., 16x16 pixels) instead processing entire images using convolutional layers detect features hierarchically. Each patch is converted into a vector and handled like "token," much like how NLP transformers handle the words. The transformer model is fed this series of patch tokens.

Positional Encoding: ViT adds positional encodings to the patch tokens in order to preserve information about their location within the original image, as transformers lacks a internal mechanism for understanding the spatial relationships between patches. This aids in an model's comprehension of the image's composition and structure.

Self-Attention Mechanism: ViT make use of an fundamental transformer feature, an self-attention mechanism, to record long-range dependencies and interactions between different patches in an image. This allows ViT to learn global context more effectively than CNNs, which focus on local features.

Vision Transformer (ViT) marks an significant innovation in applying transformer models to computer vision, providing a powerful alternative to CNNs in leveraging global self-attention mechanisms to understand visual data.

4.5.3.2 TrOCR- Transformer based Optical Character Recognition:

Text recognition is long-standing problem in research for document digitalization.This is existing approach for text recognition, are built based CNN for an image understanding and RNN for char-level text generation.

In addition, other language model is needed in improving the overall performance and accuracy as post-processing step. This is an end-to-end text recognition approach with pre-trained image Transformer and text Transformer models, namely TrOCR, which supports the Transformer architecture for both the image understanding and the wordpiece-level generation. The TrOCR model is very simple but yet effective, and can be pre-trained with large-scale synthetic data and fine-tuned with human-labeled datasets. Experiments proves that the TrOCR model outperforms the current state-of-the-art models on printed and handwritten text identification tasks.

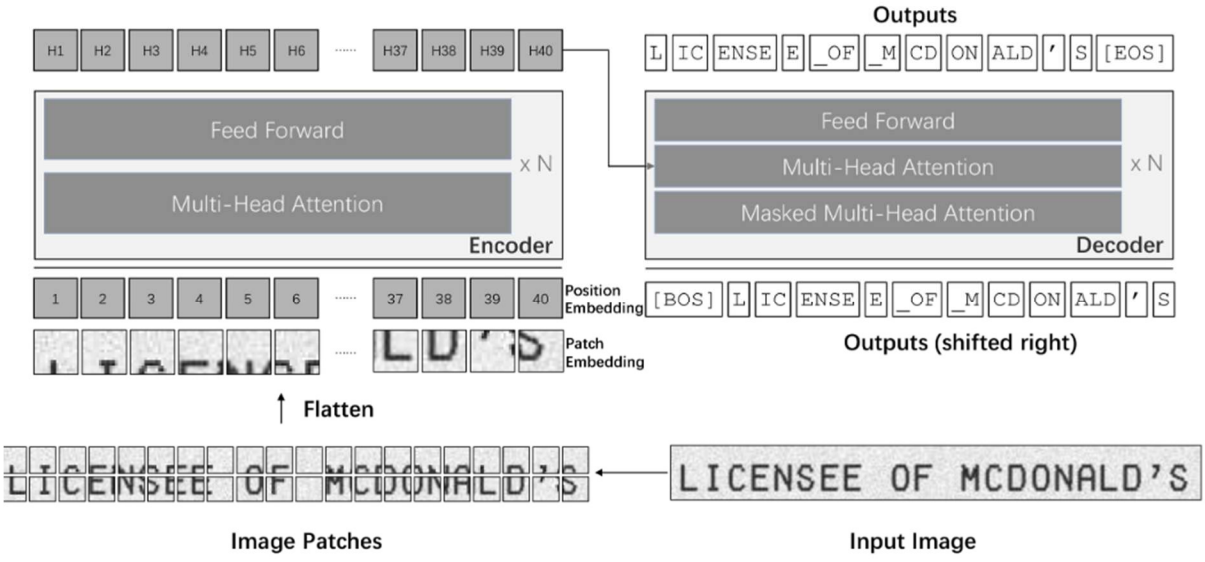

Fig.4.10 Transformer Based OCR(TrOCR)

4.6 TRAINING AND EVALUATION:

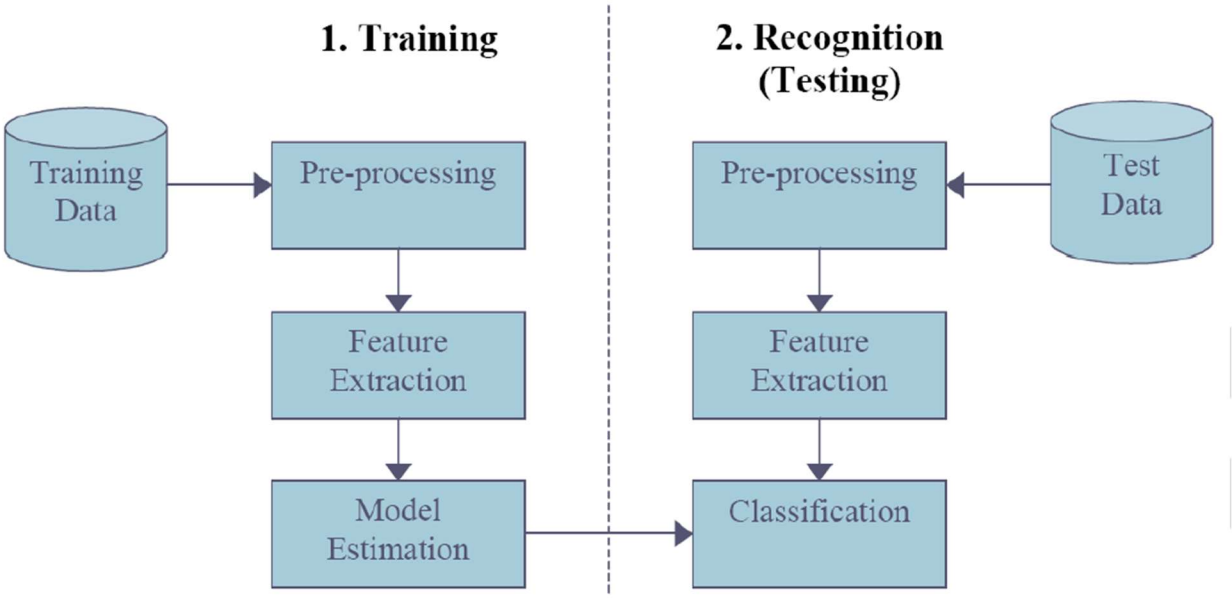

Fig.4.11 Training and Evaluation of data

The database is split into training and testing data. 80% of the data in our database is used to train and 20% to evaluate our model. The input image or picture is initially changed into a grayscale one by implementation of threshing method, that makes our image to be a binary one. This binary one is next passed into the connection test for checking the largest connecting one. Then, respective letters are further cropped to various smaller ones, which is raw-data to extract the features. These smaller images then cropped sharply into outline of letter to standardize sub-images. These images are then given as an input to the classifier. Then trained the multiple classifiers (Resnet 50, Mobilenet V2, ViT) using the training data. Then the testing data is sent to the classifier and their outputs are taken into consideration and accuracy of the model is determined. Data is also tested on the TrOCR model, which didn't yield appropriate results.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 PREPROCESSING

Input consists of a color image 3 different telugu characters.

Fig.5.1 I/P image

5.2 IMAGE ENHANCEMENT

Here the input image is the pre-processed. The colour image is then converted into grey scale image and denoising is done.

I 

Fig.5.2 Grey-Scale image

5.3 SEGMENTATION

The pre-processed image is then segmented. In segmentation, words in the image are segmented into characters using adaptive bounding box. As shown in the below figure, the bounding boxes are of various shapes depends on the size of the character. 1st bounding box is a vertical rectange and second one is horizontal.

Fig.5.3 Segmented image

5.4 CHARACTER EXTRACTION

Then, the characters are cropped from the image befor sending to the trained model. In the below figure, the 1st character from the bounding box is extracted from the image.

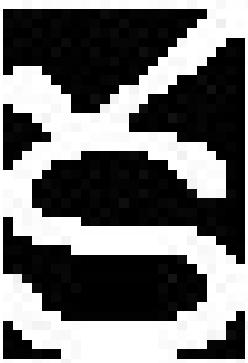

Fig.5.4 Extracted letter

5.5 ANALYZING MULTIPLE TECHNIQUES

5.5.1 PREDICTIONS OF TEST DATA ON MOBILENET V2:

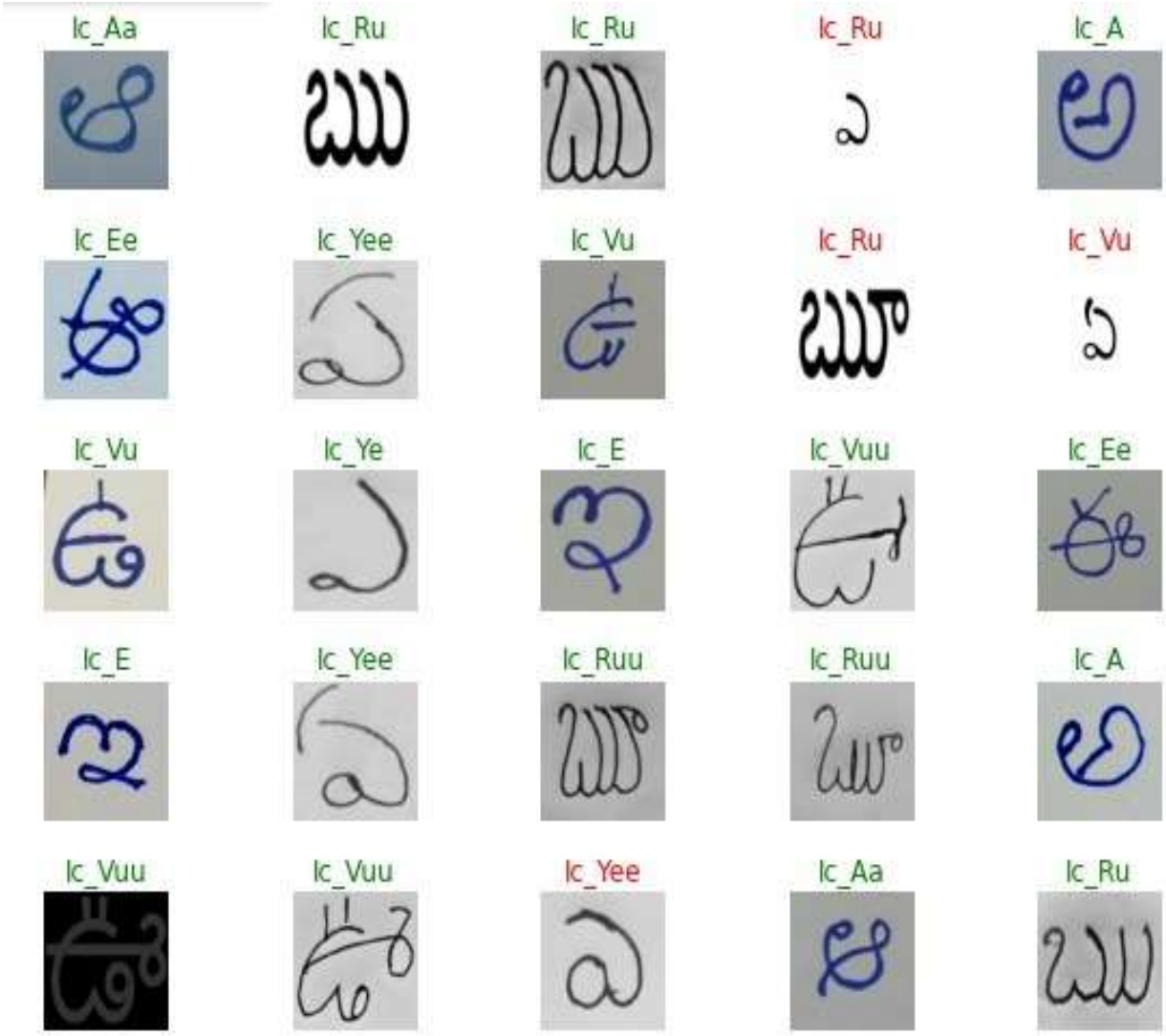

Fig.5.5 Test Dataset on Mobilenet v2

Metrics:

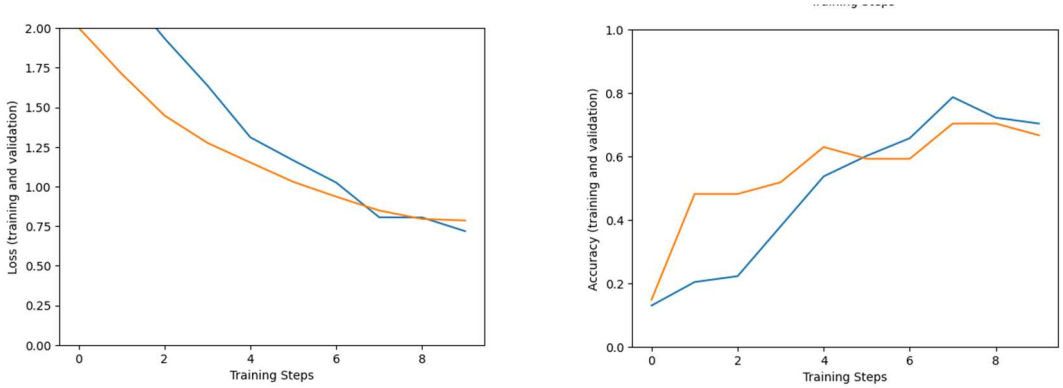

Fig.5.6 Loss and accuracy plots of Mobilenet v2

Summary: Mobilenet v2 has given 70% of accuracy which is allowed. Overall loss of Validation and accuracy drops down over the training steps, which signifies the the model is trained well. Where as the accuracy increased eventually and gave 70% of training accuracy after 5 epoches. While training the model 5 epoches are used as the data is less and using high epoch count will lead to overfitting the model. As part of the experiment, I have introduced epoch count 5, 8,10. But anything after 5 lead to overfitting. Now let's see how Resnet 50 works.

5.5.2 PREDICTIONS OF TEST DATA ON RESNET 50

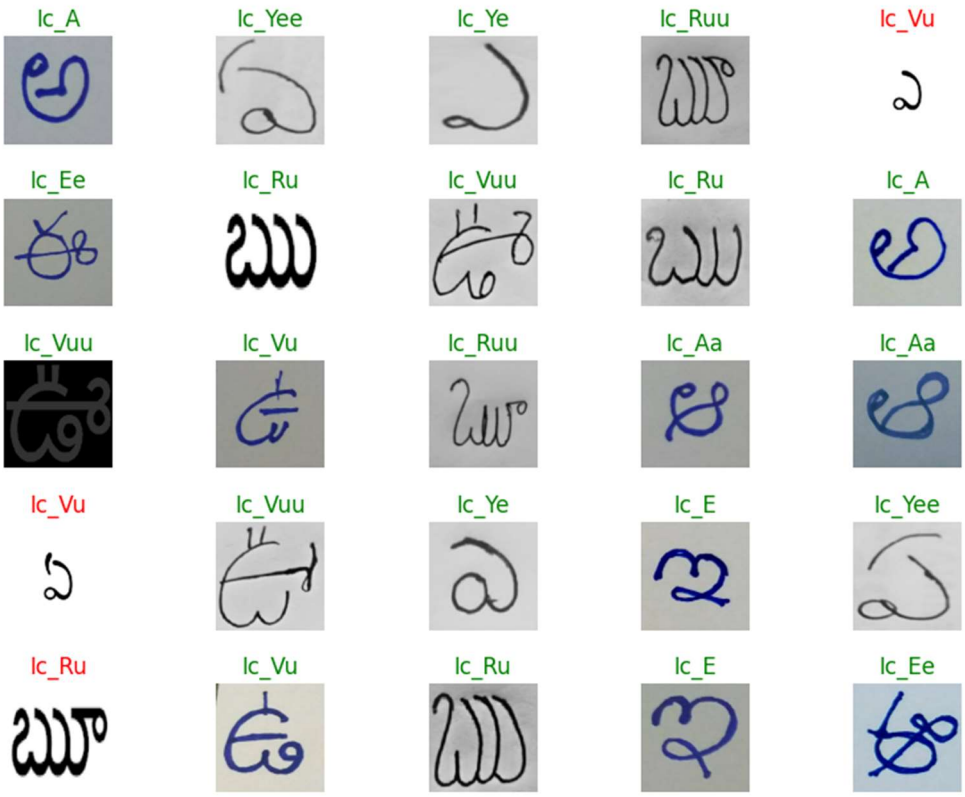

Fig.5.7 Test Dataset on Resnet 50

Metrics:

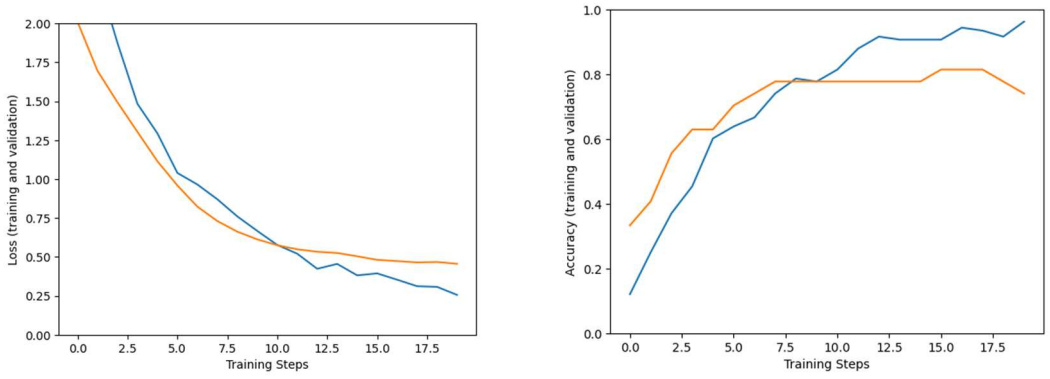

Fig.5.8 Loss and accuracy plots of Resnet 50

Summary: Mobilenet v2 has given 74% of accuracy which is allowed. Overall loss of Validation and accuracy drops down over the training steps, which signifies the the model is

trained well. Where as the accuracy increased eventually and gave 74% of training accuracy after 5 epoches. While training the model 5 epoches are used as the data is less and using high epoch count will lead to overfitting the model. As part of the experiment, I have introduced epoch count 5, 8,10. But anything after 5 lead to overfitting. Now let's see how Resnet 50 works. Resnet has gradual increase in accuracy unlike Mobilenet v2.

5.5.3 PREDICTIONS OF TEST DATA ON ViT(VISION TRANSFORMERS)

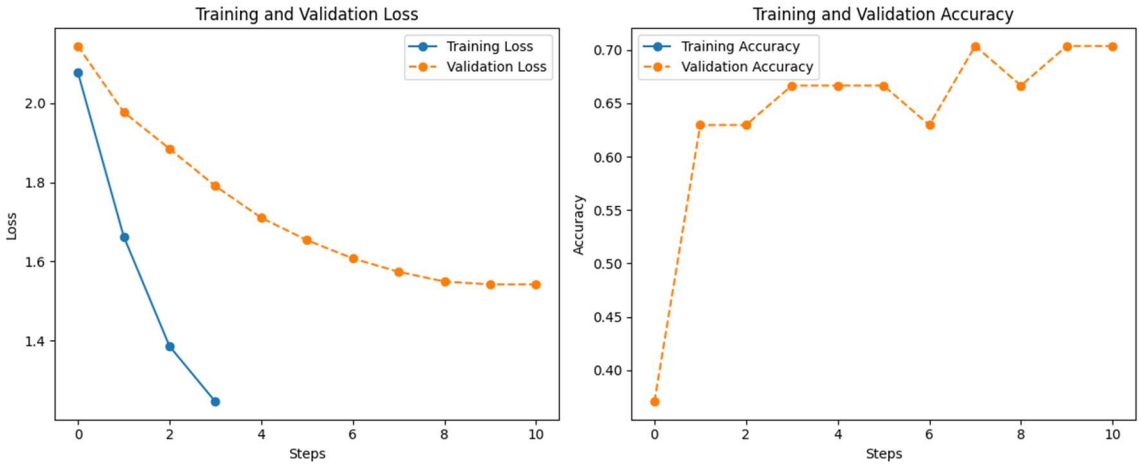

Fig.5.9 Loss and accuracy plots of ViT

Summary: The training of ViT model has been inconsistent and it didn't yield appropriate training accuracy. I have tried with hyper parameter tuning by changing the epochs count, updating the configuration, logging steps to 10, and by adding metrics callback method to fetch the training accuracy results. But, since the dataset is small, Vision Transformer did not work as expected. Data Collation and Computation metrics are also used to get the efficient results.

5.5.3.1 DRAWBACKS OF ViT(VISION TRANSFORMERS)

High Data Requirements: ViTs typically require a substantial amount of training data to achieve optimal performance.

Overfitting: With a small dataset, ViTs are prone to overfitting. Their large number of parameters and complex architectures can easily memorize the limited training examples rather than learning generalizable patterns. This results in poor performance on unseen data.

Computational Complexity: ViTs are computationally intensive, both in terms of memory and processing power. **Training them on small datasets might not leverage their full potential and can be inefficient.**

Lack of Pre-trained Models: Unlike convolutional neural networks (CNNs), which have well-established pre-trained models available for fine-tuning on smaller datasets, pre-trained ViTs are less common. This makes transfer learning less straightforward, as there are fewer pre-trained models for specific tasks and domains

Complexity of Hyperparameter Tuning: ViTs have several hyperparameters that need to be tuned carefully. On small datasets, finding the optimal hyperparameters can be challenging and time-consuming, and the model's performance can be highly sensitive to these choices.

5.5.4 PREDICTIONS ON TrOCR(TRANSFORMER BASED OCR)

A pre-trained TrOCR model is used to extract the patterns like any OCR technique performs.

It is used to convert images of text into machine-readable text. Utilizes a transformer architecture tailored for sequence-to-sequence tasks. It often includes a vision encoder (to process the image) and a text decoder (to generate text).

```python
from transformers import TrOCRProcessor, VisionEncoderDecoderModel
from PIL import Image
import torch
import matplotlib.pyplot as plt

# Load the processor and model
processor = TrOCRProcessor.from_pretrained('microsoft/trocr-base-printed')
model = VisionEncoderDecoderModel.from_pretrained('microsoft/trocr-base-printed')

# Function to predict text from an image
def predict_text(image):
    # Preprocess the image
    pixel_values = processor(images=image, return_tensors="pt").pixel_values

    # Generate the predictions
    generated_ids = model.generate(pixel_values)

    # Decode the predictions to text
    generated_text = processor.batch_decode(generated_ids, skip_special_tokens=True)[0]

    return generated_text

# Visualization of predictions
def visualize_predictions(images):
    plt.figure(figsize=(10, 9))
    plt.subplots_adjust(hspace=0.5)
    for n, image in enumerate(images):
        plt.subplot(6, 5, n + 1)
        plt.imshow(image)
        predicted_text = predict_text(image)
        plt.title(predicted_text)
        plt.axis('off')

    _ = plt.suptitle("TrOCR Model Predictions")
    plt.show()

# Example usage
image_paths = ['/content/drive/MyDrive/characters/Ic_a/1.png', '/content/drive/MyDrive/characters/Ic_ye/11.png']  # Replace with your image paths
images = [Image.open(image_path) for image_path in image_paths]

# Visualize the predictions
visualize_predictions(images)
```

Summary: The model has predicted wrong values and didn't perform as expected. This can be avoided by training a new model with required dataset.

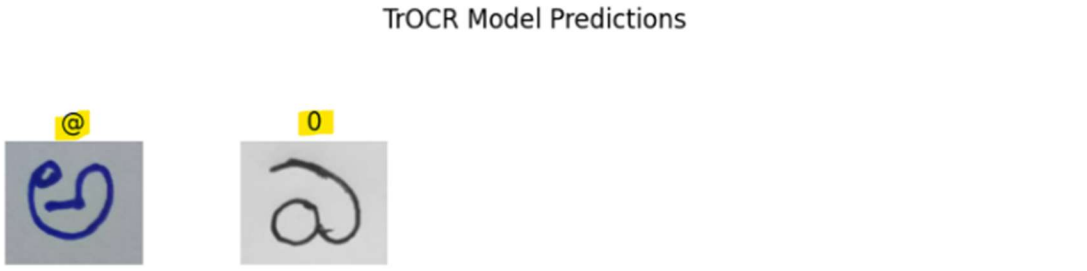

Fig.5.10 Result of Pre-trained TrOCR model

CHAPTER 6

CONCLUSIONS & RECOMMENDATIONS

6.1 Introduction

In this project, an Optical Character Recognition methodology for ancient Telugu inscriptions in temples is presented. OCR is a challenging task as various font faces and sizes exists. So, in this project I tried to compare and analyze multiple advanced OCR techniques for the character recognition. I applied image enhancement and segmentation techniques for improving the recognition accuracy of our model with the help of the reduction of noise of image. The proposed methodology can inspire to develop more Telugu OCR's. A Database of images of Telugu characters is extracted from a public dataset. This Database consists of 100 samples of 10 different Telugu characters, which is trained and tested using Convolutional Neural Network Classifier(CNN)- Resnet 50, Mobilenet v2, Transformer Based OCR and Vision Transformers. The performance accuracy of this model is achieved as **74% with Resnet 50**. By taking more samples of each character, the accuracy of our model can be improvised. This proposed work can be extended further to various other languages in India, with the scope of having a common OCR system for all the languages.

Though the advanced Transformer based are used for the research problem, they couldn't perform well due to below Drawbacks:

6.2 Limitations:

Various challenges were experienced while working on the research. Majorly related to Transformer based techniques. Assumptions while starting the research was that the Transformer based techniques would yield the best results. But, unfortunately we couldn't achieve efficient results from the recent advanced techniques. When analyzed on the cause of this issue, I could find few points.

A. **Data Requirements**:

- High-Quality Datasets: TrOCR, ViT requires large, high-quality datasets of image-text pairs for effective training. Poor or limited data can lead to suboptimal performance. The dataset we considered is less in this research problem, could be one of the limitations of not getting the appropriate results.

B. **Computational Resources:**

- Training Complexity: Training TrOCR models is computationally intensive, requiring significant GPU/TPU resources. This can be a barrier for those with limited hardware capabilities.

C. **Generalization**:

- Adaptation to New Languages or Scripts: TrOCR models trained on specific languages or scripts may need retraining or fine-tuning to handle new languages or scripts effectively. In this research, I have used Telugu language, but ViT and TrOCR are trained on English Launguage. So this can be a cause of their underperformance.

6.3 Recommendations:

Both TrOCR and ViT are powerful models with specific strengths but come with inherent limitations. TrOCR's main challenges are related to data quality and computational resources, while ViT's limitations revolve around data efficiency and computational cost. Addressing these limitations often involves leveraging large datasets, utilizing powerful computing resources, and employing fine-tuning strategies to adapt models to specific tasks or environments.

In the future research, I would recommend to use further hyper parameter tuning techniques in the recent advanced techniques and I would recommend to use high volume datasets to overcome the limitations that we have got in this thesis.

6.4 Summary:

In this chapter, we validate the research methodology previously outlined and examine the limitations encountered during the execution of this study. We also discuss the study's limitations and consider its impact and contribution to the community. The chapter concludes with suggestions for future work and opportunities for further research in this area

REFERENCES

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, International Conference on Learning Representations (ICLR), 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., “Training data-efficient image transformers & distillation through attention”, International Conference on Machine Learning (ICML), 2021.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, IEEE International Conference on Computer Vision (ICCV), 2021.

Chen, W., Xie, X., Lu, M., Wu, H., Zheng, L., Liu, C., “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation”, arXiv preprint arXiv:2102.04306, 2021.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F., Feng, J., Yan, S., “Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet”, IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M., “LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference”, arXiv preprint arXiv:2104.01136, 2021.

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S., “Rethinking Spatial Dimensions of Vision Transformers”, IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Wang, W., Cao, Y., Zhang, J., Zhang, Y., Shen, C., “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions”, IEEE International Conference on Computer Vision (ICCV), 2021.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers”, arXiv preprint arXiv:2105.15203, 2021.

Zhang, L., Chen, S., Tan, Z., Sun, L., “Multi-scale Vision Longformer: A New Vision Transformer for High-resolution Image Encoding”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., “End-to-End Object Detection with Transformers”, European Conference on Computer Vision (ECCV), 2020.

Wang, W., Zhang, J., Cao, Y., Chen, J., Xie, J., Shen, C., “End-to-End Video Instance Segmentation with Transformers”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Cai, H., Gan, C., Wang, T., Zhang, Z., Han, S., “Once-for-All: Train One Network and Specialize it for Efficient Deployment”, International Conference on Learning Representations (ICLR), 2020.

Li, Y., Yuan, Y., Huang, L., Yang, F., Guo, G., Zhang, C., Chen, X., “Enhanced Vision Transformer for Video Classification”, arXiv preprint arXiv:2106.15624, 2021.

Hu, X., Zheng, Y., Xiao, Z., Guo, Z., Wei, Y., “Pyramid Transformer for Person Re-Identification”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Pan, Z., Zhang, S., Dai, Y., Xia, C., He, X., “Scalable Vision Transformers with Hierarchical Pooling”, arXiv preprint arXiv:2104.05707, 2021.

Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X., “Deformable DETR: Deformable Transformers for End-to-End Object Detection”, International Conference on Learning Representations (ICLR), 2021.

Zhao, J., Dong, L., Xu, C., Xu, H., Zeng, M., “GridMask Data Augmentation for Object Detection”, arXiv preprint arXiv:2001.04086, 2020.

Liu, S., Hu, C., Zhang, X., Wang, X., Luo, P., “Noisy Label Detection for Large-Scale Weakly-Supervised Object Detection”, IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Lu, Y., Yuan, L., Zhang, L., Tay, F. E. H., Feng, J., “Pre-training Image Transformers with Large-scale Jigsaw Puzzles”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., Patel, V. M., “Medical Transformer: Gated Axial-Attention for Medical Image Segmentation”, IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Gao, P., Zheng, M., Wang, X., Dai, J., Li, H., “Fast Vision Transformers with Hierarchical Attention”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Kumar, S., Varshney, H., Khan, S. M., Mitra, S., “ViT-Dehazing: A Vision Transformer for Single Image Dehazing”, IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., “Transformer in Transformer”, Advances in Neural Information Processing Systems (NeurIPS), 2021.

Zhang, H., Li, C., Han, K., Wang, Y., Tian, Q., “Multi-scale Vision Transformers with Dynamic Token Pooling”, arXiv preprint arXiv:2104.11734, 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., “Augmenting Convolutional networks with attention-based transformers”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Wang, H., Wang, T., Zhang, Z., Han, Y., “Transformer-based Convolutional Neural Networks for High-Resolution Aerial Imagery Segmentation”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Wang, W., Cao, Y., Zhang, J., Yu, Z., Xie, J., Shen, C., “CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Yu, S., Fang, Z., Zhang, X., “Visual Prompting for Adapting Large-Scale Models to Vision Tasks”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., “Scaling Vision Transformers”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Vamvakas, G. & Gatos, B. & Stamatopoulos, Nikolaos & Perantonis, Stavros, “A Complete Optical Character Recognition Methodology for Historical Documents”, Document Analysis Systems, IAPR International Workshop on. 525-532. 10.1109/DAS.2008.73.

Pratik Madhukar Manwatar and Shashank H.Yadav, “Text Recognition from images”, International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015

Haifeng Zhao, Yong Hu and Jinxia Zhang, ”Character Recognition via a Compact Convolutional Neural Network”, International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2017

Shreshtha Garg, Kapil Kumar Gupta and Nikhil Prabhakar, “Optical Character Recognition using Artificial Intelligence”, International Journal of Computer Applications (0975 – 8887), Vol.179, No.31,April 2018

Aravinda C.V, Lin Meng and Udaya Kumar Reddy, “A Complete Methodology for Kuzushiji Historical Character Recognition using Multiple Features Approach and Deep Learning Model”,International Journal of Advanced Computer Science and Applications, Vol.11, No.8, 2020

Prof. Sheetal A. Nirve and Dr. G. S. Sable, “Optical character recognition for printed text in Devanagari using ANFIS”, International Journal of Scientific & Engineering Research, Vol.4, Issue 10,2013

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems

(NeurIPS). Available at: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Long, J., Shelhamer, E., and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf

He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Sutskever, I., Vinyals, O., and Le, Q. V. (2014) Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems (NeurIPS). Available at: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017) Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS). Available at: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014) Learning Deep Features for Scene Recognition using Places Database. Advances in Neural Information Processing

Systems (NeurIPS). Available at: https://papers.nips.cc/paper/5548-learning-deep-features-for-scene-recognition-using-places-database.pdf

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. International Conference on Machine Learning (ICML). Available at: https://arxiv.org/pdf/1502.03044.pdf

Keaton, M. R., Zaveri, R. J., and Doretto, G. (2023) CellTranspose: Few-Shot Domain Adaptation for Cellular Instance Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Keaton_CellTranspose_Few-Shot_Domain_Adaptation_for_Cellular_Instance_Segmentation_WACV_2023_paper.html

Yang, F., Odashima, S., Masui, S., and Jiang, S. (2023) Hard To Track Objects With Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Yang_Hard_To_Track_Objects_With_Irregular_Motions_and_Similar_Appearances_WACV_2023_paper.html

Bera, S., and Biswas, P. K. (2023) Self-Supervised Low Dose Computed Tomography Image Denoising Using Invertible Network Exploiting Inter Slice Congruence. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Bera_Self-Supervised_Low_Dose_Computed_Tomography_Image_Denoising_Using_Invertible_Network_WACV_2023_paper.html

Aich, A., Li, S., Song, C., Asif, M. S., Krishnamurthy, S. V., and Roy-Chowdhury, A. K. (2023) Leveraging Local Patch Differences in Multi-Object Scenes for Generative Adversarial Attacks. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Aich_Leveraging_Local_Patch_Differences_in_Multi-Object_Scenes_for_Generative_Adversarial_WACV_2023_paper.html

Li, M., Zhang, H., Zhang, S., Liu, X., Zhou, M., Wei, F., & Zhou, L., "TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models", arXiv preprint arXiv:2109.10282, 2021.

Fang, S., Xing, J., Zhang, L., Xie, L., "Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition", International Conference on Computer Vision (ICCV), 2021.

Baek, J., Lee, B., Han, D., Yun, S., Lee, S., "Character Region Awareness for Text Detection", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Wang, W., Xie, E., Li, X., Hou, Y., Lu, T., Yu, G., Shao, L., "PAN++: Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text", IEEE Transactions on Image Processing, 2021.

Hu, Y., Zhang, X., Li, Y., Zhang, X., "GTC: Guided Training of CTC Towards Efficient and Accurate Scene Text Recognition", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Cheng, Z., Xu, Y., Bai, F., Niu, Y., Pu, S., Zhou, S., “AON: Towards Arbitrarily-Oriented Text Recognition”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

Dai, Y., Liu, Y., Jin, L., Zhang, S., Luo, C., “Fused Text Segmentation Networks for Multi-oriented Scene Text Detection”, International Conference on Image Processing (ICIP), 2018.

Yang, M., Liu, X., Luo, J., Sun, W., “Transformer-based End-to-End Image Text Recognition with Pre-trained Models”, arXiv preprint arXiv:2103.14322, 2021.

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J., “EAST: An Efficient and Accurate Scene Text Detector”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

Shi, B., Bai, X., Yao, C., “An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

Biten, A. F., Mafla, A., Gomez, L., Zhu, M., Van Gool, L., Jawahar, C. V., “Scene Text Visual Question Answering”, IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

Liu, W., Chen, C., Shen, C., He, T., “ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Wang, Y., Xie, Z., Liao, M., Liang, J., “Towards Accurate Scene Text Detection with Semantic Segmentation as Extra Supervision”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Zhang, L., Yu, J., Hu, W., “A Hybrid Transformer Network for Scene Text Recognition”, arXiv preprint arXiv:2104.01292, 2021.

Qiao, Z., Zhang, Z., Tang, J., Wan, C., Liu, Y., “Text Perceiver: An All-in-One Transformer for Scene Text Detection and Recognition”, arXiv preprint arXiv:2108.04930, 2021.

Xiao, L., Shi, X., He, D., Lin, Z., “CoTNet: Context Transformer Networks for Scene Text Detection”, International Journal of Computer Vision (IJCV), 2021.

Zhu, Y., Tang, X., Qiao, Z., “Self-Training with Noisy Student Improves Scene Text Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.

Wan, Z., He, S., Zhang, W., “SGTR: Scene Text Recognition with Semantic Guidance and Transformer”, arXiv preprint arXiv:2106.01242, 2021.

Zhu, Y., Tang, X., Cheng, Z., “Ensemble of Adaptive Local Monitors for Scene Text Detection”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Wu, X., Liu, Y., Chen, X., “Transformer-based Dual Decoder for Scene Text Recognition”, IEEE Transactions on Multimedia, 2021.

Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A., “Reading Text in the Wild with Convolutional Neural Networks”, International Journal of Computer Vision (IJCV), 2016.

Li, H., Wang, P., Shen, C., “Towards End-to-End Text Spotting with Convolutional Recurrent Neural Networks”, IEEE/CVF International Conference on Computer Vision (ICCV), 2017.

Zhai, X., Tang, Y., Shi, B., Zhang, X., “Learning Discriminative Feature Representation for Scene Text Detection with Transformers”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Lin, T., He, D., Lin, Z., “Text Transformer: An End-to-End Transformer Network for Scene Text Recognition”, arXiv preprint arXiv:2104.01298, 2021.

Long, S., Ruan, J., Zhang, W., He, X., Wu, W., “Two-step Deep Neural Network for Text Segmentation and Recognition in Complex Scene Images”, IEEE Transactions on Image Processing, 2019.

Cho, D., Sung, J., Han, D., “Synthetically Supervised Feature Learning for Scene Text Recognition”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Tian, Z., Huang, W., He, T., Qiao, Y., “Detecting Text in Natural Image with Connectionist Text Proposal Network”, European Conference on Computer Vision (ECCV), 2016.

Chen, S., Li, X., Bai, X., “Transformer-based Text Recognition: Towards Real-time and Robust Performance”, arXiv preprint arXiv:2103.09460, 2021.

Zhou, X., Yao, C., Wen, H., Wang, Y., “Densebox: Unifying Landmark Localization with Object Detection”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Yuan, Y., Zhan, Y., Tang, X., “A Comprehensive Survey of Vision Transformer for Computer Vision Applications”, arXiv preprint arXiv:2111.12176, 2021.

Appendix:

Recognition of Telugu Ancient Characters And Information Retrieval From Temple Epigraphy Using Deep Learning

VEENA SAI NIGAMA

Research Proposal

Master of Science in Machine Learning & Artificial Intelligence

Liverpool John Moores University & upGrad

MAY 2024

Abstract

Illuminating pivotal role of ancient Hindu temples in preserving history of India carved on stone. A temple is projection of information, from the time of creation and by the people constructed it. Epigraphy is the science of deciphering the inscriptions, is the primary tool of scientists to study past, which might impact our present or future. Most of the temple inscriptions are in ancient Indian languages which are challenging for tourists to understand. Tourist guides are rare, and some ancient temples may not be accessible to locals due to outdated language. Image is captured as input and noise of image is reduced by image enhancement techniques. In this project we will try to achieve maximum efficiency and also reduce the duration of time for the character recognition. Segmentation techniques like line, word/character are used to extract main components of image. So, with the help of extracting features, important features are extracted leaving behind the undesired. Next with the help of a neural network, training, classification and recognition of the Telugu letters is done. For character recognition, python libraries are used in Telugu character database, consisting of 100 samples trained and tested using Convolutional Neural Network Classifiers (Resnet, VGG, TrOCR, ViT) . By taking more samples of each character, the accuracy of our model can be improvised. This proposed work can be extended further various languages which are spoken in India.

Contents

LIST OF FIGURES

1. Background

Our country, India, is very famous for its rich culture and heritage. One thing in which India faces no competition in its different heritage and culture. The big history of India, played a great role in shaping the culture of the country. The geography of the country is also very unique. Though our country has absorbed customs and traditions from other countries, it also preserved the heritages of ancient times. Also, India is famous for its diversity in almost everything.

For our work considered temples in South Indian states like Andhra Pradesh and Telangana. The language that is spoken commonly among the two states is Telugu. So, the proposed work mainly focuses on developing an OCR system for digitizing and producing an understandable output for Telugu inscriptions mainly discovered on the floors, walls and pillars of temples located in two states. Optical Character Recognition also known as OCR is simply recognizing handwritten, typed, printed letters by an electronic device such as computer

In this project, trying to extract the characters from the image that is taken by the user, and to convert it into a meaningful text in whichever language the user is comfortable with. Extraction and translation are two different phases after feature extraction. In order to make it user friendly, this project can be extended to the comfortable platform like creating a Web App, or normal Android App. By extending this project, the work becomes user friendly. If the user imports a picture of Inscription to the Web App or Mobile App, the model that was trained gives the meaningful output text to the user in the language he/she is comfortable with. This project not only adhere to Inscriptions, it can be used to get the meaning of any telugu text, as the model trained the model with Handwritten characters.

This project is going to analyze multiple OCR techniques (Resnet, VGG, TrOCR, ViT-Vision Transformers) in order to achieve efficient model as part of this work.

2.Problem Statement & Related Work

2.1 Problem Statement

The main aim of this research is to develop the robust system for extraction and recognizing telugu handwritten characters from images of ancient temple inscriptions. Using advanced neural network techniques like CNNs, TrOCR(Transformer based OCR), ViTs(Vision transformers), the goal is to accurately digitize and archive these inscriptions, preserving valuable historical, cultural, and linguistic information.

2.2 Related work

In the paper, “Optical Character Recognition using Convolutional Neural networks”, by Sakshi Shreya, Gagan Upadhyay, Mohit Manchand, Rubeena Vohra and Gagan Deep Singh, firstly a picture is scanned, then analysis of the scanned image and finally the images of characters are translated into their corresponding ASCII codes. Segmentation algorithm is used and a convolutional neural network is used for recognition as the results given by CNN are more accurate when compared to other neural networks and machine learning algorithms.

In the paper, “Character recognition via a Compact Convolutional Neural network”, by Haifeng Zhao, Yong Hu and Jinxia Zhang, optical character recognition of images from natural scenes is done using a type of CNN called VGG- Net. They focused on compacting the architecture of the neural network so that both word and character segmentation can be done under the same framework thus reducing the complexity.

In the paper, “A Generic OCR Using Deep Siamese Convolutional Neural networks”, by Ghadha Sokar, Elsayed E. Hemayed and Mohamed Rehan, optical character recognition is done by using deep siamese CNN and Support Vector Machines. Several problems like fine-tuning a trained model are addressed by this paper. The model is trained for extracting the distinct features of the characters.

In the paper, “ Text Recognition from Images” by Pratik Madhukar Manwatkar and Shashank H. Yadav, text recognition is divided into four modules namely ; pre-processing, system training, text recognition, and post processing. Matrix feature extraction method is used. An Artificial Neural Network called Kohonen Neural Network is used as it has the capability to train itself automatically.

After studying various papers and knowing about the various techniques used for optical character recognition, it is decided the workflow for the project. The image is taken as the input by reading or scanning and it is converted as image of the digital type. Enhancement techniques like binarization can be helpful for reducing noise in the image. Segmentation technique like line, word or character segmentation are used to extract the main components of image. With the help of extracting features, important features are extracted leaving behind the undesired. Then the neural network is used for training, classification, and recognition of old Telugu letters.

2. Aim and Objectives

The primary aim of this research is to analyze and propose the efficient technique to perform OCR with better performance. The identification of the Telugu handwritten character using Neutral Networks.

The below research objectives are formulated based on the aim of the study

- To conduct an analysis in order to find a technique in terms of performance and results.
- To explore the viability and then develop a balancing technique which will obtain efficient results and performance.
- To evaluate the performance of the best performing model's accuracy.

4. Significance of the Study

OCR being the active research filed, vigorous research is happening over the world currently. There is a gap in analysing the techniques based on performance.

This research fills in gaps through adding to existing literature, by contributing to code. The work explores advanced developments in recent times in Transformers based OCRs.

5. Scope of the Study

Below is the scope of thesis:

- The research work to be finished within 15 weeks post to submitting research proposal report.
- This evaluation and analysis to be conducted through open source modules & models.
- The model training and classification to be conducted with available public GPU.

6. Research Methodology

The project will involve acquiring high-resolution images of temple inscriptions and enhancing their quality through preprocessing techniques such as noise reduction and binarization. Characters will be isolated using segmentation methods. Convolutional neural networks (CNNs), and Transformer based OCR & Computer Vision models will be employed to recognize the handwritten characters.

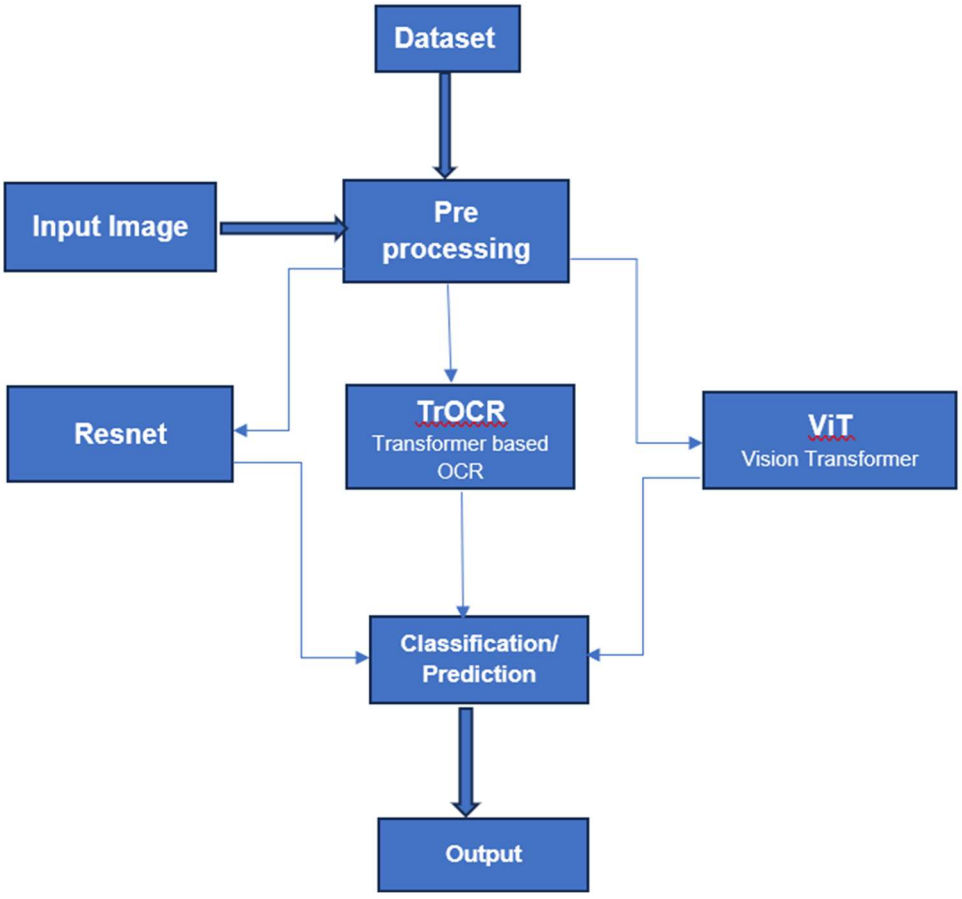

Figure 6.1

6.1. Dataset description:

Dataset consists of all Telugu handwritten characters taken from below IEEE site.

https://ieee-dataport.org/open-access/telugu-handwritten-character-dataset

File name of each character is embedded with its ‘English’ transliterated form.

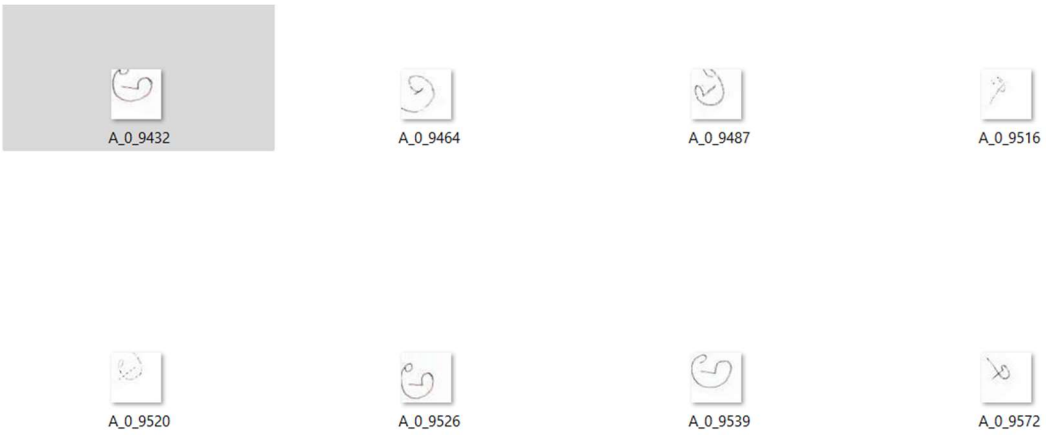

Figure 6.11

6.2. Data preprocessing:

- Apply preprocessing techniques such as Gaussian blurring and Otsu's thresholding to enhance image quality.

- Use morphological operations and edge detection to segment characters from the background.

6.3. Algorithms & Techniques:

- Implement CNNs for effective character recognition from images.

- Explore transformer models like TrOCR, ViT for advanced recognition and context understanding.

7. Requirements Resources

7.1 Hardware Requirements

The below hardware requirements must be met for the research work as follows:

- Laptop or a desktop along with proper internet which is capable of browsing & document writing and executing code.
- To execute the CUDA based neural networks, GPU is required.

7.2 Software Requirements

The below software requirement must be met for the research work:

- IDE
- Deep Learning libraries - TensorFlow, HuggingFace and PyTorch
- Python 3.7+
- Data analysis libraries- NLTK, Pandas, Numpy, etc.

8. Research Plan

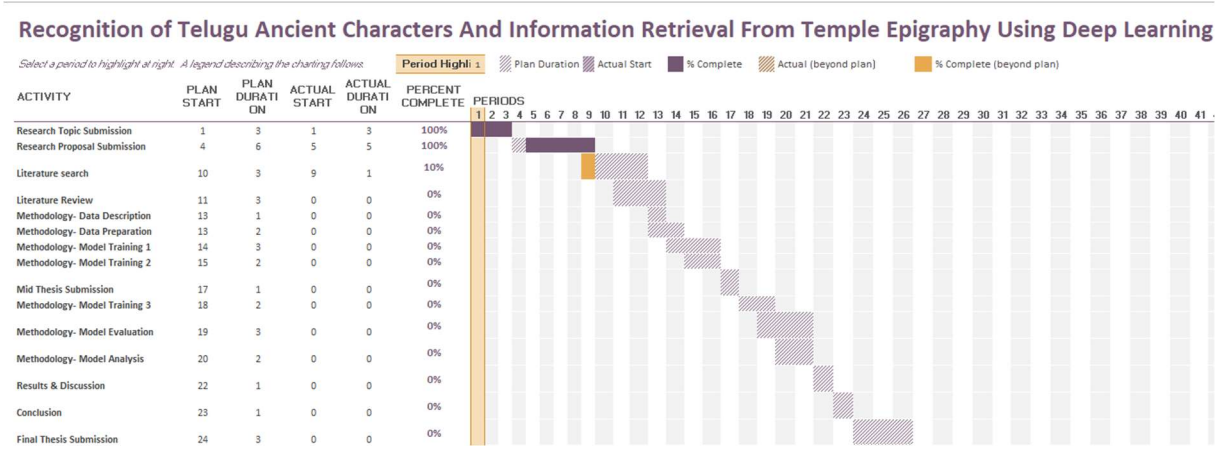

Figure 8.1

9. References

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*. Available at: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Long, J., Shelhamer, E., and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf

He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Sutskever, I., Vinyals, O., and Le, Q. V. (2014) Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*. Available at: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*. Available at: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014) Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems (NeurIPS)*. Available at: https://papers.nips.cc/paper/5548-learning-deep-features-for-scene-recognition-using-places-database.pdf

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International Conference on Machine Learning (ICML)*. Available at: https://arxiv.org/pdf/1502.03044.pdf

Keaton, M. R., Zaveri, R. J., and Doretto, G. (2023) CellTranspose: Few-Shot Domain Adaptation for Cellular Instance Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Keaton_CellTranspose_Few-Shot_Domain_Adaptation_for_Cellular_Instance_Segmentation_WACV_2023_paper.html

Yang, F., Odashima, S., Masui, S., and Jiang, S. (2023) Hard To Track Objects With Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Yang_Hard_To_Track_Objects_With_Irregular_Motions_and_Similar_Appearances_WACV_2023_paper.html

Bera, S., and Biswas, P. K. (2023) Self-Supervised Low Dose Computed Tomography Image Denoising Using Invertible Network Exploiting Inter Slice Congruence. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Bera_Self-Supervised_Low_Dose_Computed_Tomography_Image_Denoising_Using_Invertible_Network_WACV_2023_paper.html

Aich, A., Li, S., Song, C., Asif, M. S., Krishnamurthy, S. V., and Roy-Chowdhury, A. K. (2023) Leveraging Local Patch Differences in Multi-Object Scenes for Generative Adversarial Attacks. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Available at: https://openaccess.thecvf.com/content/WACV2023/html/Aich_Leveraging_Local_Patch_Differences_in_Multi-Object_Scenes_for_Generative_Adversarial_WACV_2023_paper.html

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. (2023) Magic3D: High-Resolution Text-to-3D Content Creation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://arxiv.org/abs/2303.04534

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. (2023) Align Your Latents: High-Resolution Video Synthesis With Latent Diffusion Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://arxiv.org/abs/2303.04535

Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., and Williams, F. (2023) Neural Kernel Surface Reconstruction. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://arxiv.org/abs/2303.04536

Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., and Liu, S. (2023) Affordance Diffusion: Synthesizing Hand-Object Interactions. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://arxiv.org/abs/2303.04537

Li, Z., Huang, J., Gojcic, Z., Litany, O., and Fidler, S. (2023) Neuralangelo: High-Fidelity Neural Surface Reconstruction. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://arxiv.org/abs/2303.04538

Liu, R., and Vondrick, C. (2023) Humans As Light Bulbs: 3D Human Reconstruction From Thermal Reflection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://arxiv.org/abs/2303.04539