Language Interpreter and Speaker

Ruchi Bari
Information Technology
Vidyavardhini's College of Engineering and Technology
Vasai Road(W), India
ruchi.bari85@gmail.com

Mrunmayee Apte
Information Technology
Vidyavardhini's College of Engineering and Technology
Vasai Road(W), India
mrunmayeeapte27@gmail.com

Aakanksha Mohite
Information Technology
Vidyavardhini's College of Engineering and Technology
Vasai Road(W), India
aakanksharmohite@gmail.com

Sainath Patil
Dept. of IT, VCET
University of Mumbai
Mumbai, India
sainath.patil@vcet.edu.in

***Abstract*—Language Interpreter and Speaker is a device for identifying the language of the written image text and then converting the same text to speech format. This device would surely be useful for blind and visually impaired people. Language identification (LI) is the method in which we identify the natural language of the given content. It is the process of categorizing a document on the basis of its language. In this developing generation, we are heading towards a phase where computers would be capable of doing all such things that humans can do. Recognition of language used is the initial requirement before reading or learning. To start with any of the tasks, humans first try to understand the task and then process the task. Similarly for language identification, the machine needs to learn the language and then once learning is completed it should be able to recognize the language. The project is divided into three parts. Initially, the handwritten image text would be converted to the normal text. In the second part the language would be identified from the converted text and lastly, the text would be converted to audio format. This paper will discuss the implementation of this idea, give an approach to problems and challenges that we came across, and some solutions.**

***Index Terms*—Image Processing, CNN (Convolution Neural Network), AlexNet, gTTS(google-text-to-speech)**

I. Introduction

A. Problem Statement

In today's world, language and speech are very important for communicating with each other. Due to the evolution of technology, there are a number of traditional ways, social media platforms, etc. for humans to gossip, share ideas, and knowledge with each other. Knowing a language is not only important for conversation, it also helps to build relationships and a sense of community among the people. For now, there are around 6900 languages spoken all around us and each of them is unique in a number of ways. There would be no such person on the earth who may be knowing all the existing languages. Thus, knowing the language of the text is the primary step of any type of communication. This paper presents the outline done for predicting the language from handwritten as well as digital image text.

B. Motivation

At times we humans get bored of reading large content text or even sometimes due to workload we get exasperated to read long content text by ourselves. Every individual has faced this situation at some instant. To overcome this problem, this paper gives a solution of transforming the same user input text into audio format. Also, this conversion of text into speech format would not only be useful for normal people but would also be helpful for visually impaired people. As we know visually challenged people are mostly dependent on others for getting information from digital images or handwritten documents. In order to assist them, Braille is one of the systems that was developed. However the main drawback of this system is that it absorbs more time. Also, we cannot use this method on digital images. So the text-to-speech converter would be the easiest and most helpful method for them. The project's purpose is to learn how to make this tedious job easy by using CNN by constructing a Language Interpreter and Speaker tool that shows the language of the user input image text and also the audio format of the same text. First, the entered image undergoes the process of image processing. After that by using the CNN model is trained then the language is predicted and by using gTTS it is converted into a speech format.

II. REVIEW OF LITERATURE

We produced our project Language Interpreter and Speaker with the aid of all of this following paper -

The authors of [1] designed a system using OCR 'Optical Character Recognition'. Pre-processing, segmentation, feature extraction, and post-processing this are some important steps which are followed by OCR. With the help of this system we can easily edit or share the recognized data having 90 percent accuracy for handwritten documents. Using an Android app, this approach converts Handwritten Character Recognition into editable text. The camera captures an image and

saves it to the PC.The user is given the choice of picking a component using an android app of a file that has to be converted Further processing is required. The OCR engine helps to convert the text and displays it on the screen. The text that has been analyzed is kept in text format to make changes to the text that has been recognised a choice is made, and they must be stored proper location. When text is printed, it is more accurate rather than in handwritten form.

The writers of [2] have proposed a technique to detect the language of a text document image that contains Indian languages in this paper. Here they used India's 3 major languages English, Hindi, and Tamil. They trained a CNN model on images of individual characters of each language. Around 13000 images each of English, Hindi, and Tamil were taken. The network is trained on characters from various languages, and accuracy of roughly 74 percent was reported. Their model has three output nodes giving three different probabilities for languages. The language showing the highest probability is considered the language of that input text.

The authors of [3] came up with the idea of a smart reader has been suggested to help society's outwardly challenged individuals by detecting text, recognising the face of a familiar person, translating the identified word, and producing speech output. This can assist the user in reading any material, recognising a recognised person, and receiving the result in vocal form.

With the aid of CNN, the developer of [4] detailed handwritten character recognition from photos. Optical Character Recognition (OCR) uses an optical picture of a character as an input and outputs the corresponding character. They demonstrated OCR using a CNN and Error Correcting Output Code (ECOC) classifier combination. The CNN is used to extract features, while the ECOC is used to classify them.

Amit Choudhary [5] first performed binarization technique on image which was followed by feature extraction for offline Handwritten Character Recognition. Binarization helps to extract features of hnadwritten english characters. The algorithm used gave classification accuracy of 85.62 percent.

For this project datasets were obtained from Kaggle Handwritten Character [6], IAM-Dataset [7]. The authors of [2] created their own dataset after collecting images form different websites. Also there are different types of convolutional neural networks which are used for OCR. In paper [4] authors have performed research on accuracy of different CNN models such as AlexNet, Lenet, Zfnet, ect. From which Alexnet gave highest accuracy among all those models.

III. Method

The project is mainly divided into three parts. In this section we present our procedure for making the project, describing all three phases of the project. The procedure is divided into smaller steps to attain maximum accuracy, success, and is free of bugs. Following is the description of all the three parts of the project:
The entire project is divided into three steps as :
1. Handwritten/Digital text recognition
2. Finding the language of the text
3. Converting text into speech

1. **Handwritten text recognition**:
The primary step of our project is to extract text from an image having handwritten text. This step is a very important step of our project as it will be input for the other two parts of the project.The next crucial part of the project is to train and test the model. For training our model we have made use of CNN - Alexnet model. We have made two sections to the dataset: training and validation. There are 140000 images in training and 15209 images in validation.

2. **Finding the language of the text**:
The goal of this section of the research is to predict the language of the detected image text that was obtained in the first part. For language identification, we have made use of stopwords. By using stopwords the language of the image text is displayed on the screen.

3. **Conversion of text into speech**:
This is the third part of our project where the image text is transformed in speech format. To achieve this we have made use of the gTTS API. By using this the text is converted to audio format.

The closing step of the project is the UI/UX. For UI, Streamlit is used. It is nothing but a framework to construct web apps for ML and Data Science. Python is a programming language. It's a straightforward intermediary between the model and the web app.

IV. Solution Approach

Working
As discussed in the overview the project is divided into three steps. This section gives information about the detailed work done.

1. Handwritten text recognition
To convert text from handwritten text image to digital text this is the list of steps that were followed:

1.1 Image gathering:
The system takes an image taken using a mobile phone and

executes image processing procedures on it. This system will also crop the extra white spaces around the image which helps the contouring better.

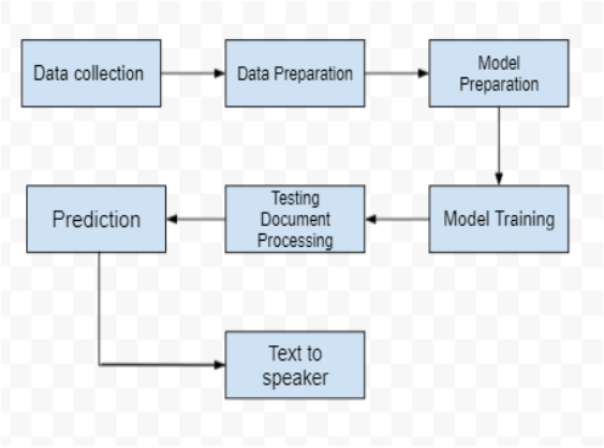

Fig. 1. Flowchart of the process

1.2 Preprocessing and segmentation:
Because a handwritten text image is more sensitive to noise, preprocessing is one of the most critical phases in text recognition. It removes impurities from the image to make the image more readable for the computer. Then in the next step segmentation is performed in which each character of the word is separated to predict more accurately. First, the line is broken down into words and then each character from the word is recognized separately. For user input we take an image clicked by a phone from which we first crop the extra white spaces. Then the line is broken down in words and then each character from the word is recognized.

1.3 Feature extraction:
The procedure in which OCR recognises alphabets based on distinct classes is called feature extraction. The translation of input data into a set of features is known as feature extraction. The features are extracted from the text image. Their qualities are their features. Slant angle, height, curves, and other factors are used to classify the alphabets. The selected text is compared to the system's standard database and the dataset, with the strongest correlation being chosen and defined as a character. The depiction of symbols is the focus of feature extraction. The character is turned into text once it has been recognised based on classification. Feature detection provides information about the characteristics of numbers or letters on an individual basis, allowing characters in a document to be recognised.

1.4 Contours:
They are the lines that connect all of the locations along the image's boundaries that have the same intensity. In this project, we are using contours for converting the handwritten image text to digital text. A boundary is drawn around the letter and that contoured image is used for the prediction of the letter that matches with the image from the datasets.

1.5 Post-processing:
The computer is the only one who understands the extracted output. As a result, data must be saved in a specific format (.txt). The ASCII data was created from the recognised data. The following are the steps involved in this work are -
1. Image acquisition using an Android camera
2. Loading the image on the Android Studio-created Graphical User Interface (GUI).
3. Preprocessing of the image.
4. Feature extraction from the input image
5. OCR is used to turn recognised data into text format

2. Finding the language of the text
After the conversion of handwritten or digital image text to digital text, the next step is the language identification part. This fragment of the project displays the language identified on the output screen. For doing this we have made use of stopwords.

2.1 Stopwords:
These are the list of words that are more commonly used words in a language. They have very little meaning, but are often used. Stop words in English include ”the,” ”is,” ”are,” ”in,” and so on. All these types of words can be used to identify the language of the digitally converted text

2.2 Tokenization:
It Is a process of splitting a sentence or paragraph into small units called tokens. This is an important step if we want to get the meaning of the sentence given, as the words present in the sentence give us the meaning of the sentence rather than considering a whole sentence. For example, “Technology is Good” can be tokenized into [“Technology”, “is”, “Good”]. This helps us to determine the number of words in the sentence, it can also help us get information about the frequency of a particular word in the sentence.

2.3 Training and Testing:
For the aim of training there are many classification algorithms among which we have used the AlexNet model of CNN algorithm as it gave high accuracy when it came to text analysis.

Alexnet
AlexNet is one of the models of CNN. CNN is a sort of artificial neural network that is specifically built to process pixel data and is used in image recognition and image processing. CNN popularised AlexNet, a deep learning architecture. Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton created it. It has similar architecture as that of LeNet but AlexNet is deeper and has more filters.

The Alexnet is made up of eight layers, each having its

Layer	Filters	Filter size	Stride	Padding	Size of feature map	Activation function
Input	-	-	-	-	227x227x3	-
Conv1	96	11x11	4	-	55x55x96	ReLU
Max Pool1	-	3x3	2	-	27x27x96	-
Conv2	256	5x5	1	2	27x27x256	ReLU
Max Pool2	-	3x3	2	-	13x13x256	-
Conv5	256	2x2	1	1	13x13x256	ReLU
Max Pool3	-	1x1	2	-	6x6x256	-
Dropout1	rate=0.25	-	-	-	6x6x256	-

TABLE I
ARCHITECTURE OF ALEXNET

own set of parameters that may be learned. The model is made up of five convolution layers with a max pooling combination, followed by three fully connected layers. Except for the output layer, each of these levels uses the Relu activation function. 1000 neurons make up the final completely linked layer, often known as the output layer. Softmax is the activation function used in the last layer of this model.

Following are some of the models that we tried on the dataset:-

Architecture	Training accuracy	Validation accuracy
Alexnet-1	0.9605	0.9095
Alexnet-2	0.9827	0.9118
Lenet	0.9066	0.9032
Lenet-5	0.9567	0.8907
Custom-Model	0.9509	0.9172

TABLE II
DIFFERENT USED MODELS ACCURACY

3. Converting text into speech

The conversion of text to speech is the process of converting words into vocal audio format. In the traditional method, the text from handwritten text images will be extracted by the application, which will then be analysed using natural language processing and digital signal processing.
gTTS i.e. Google Text-To-Speech (gTTS) is a Python library. This is a command-line interface for interacting with the Google Translate Text-to-Speech API. The basic requirement for using the gTTS library is one must have a version of Python greater than 2.7.

V. RESULT AND DECISION

The result of implementing the language of text and audio file of a specified text on a user input image.
Following are some of the implementation results:

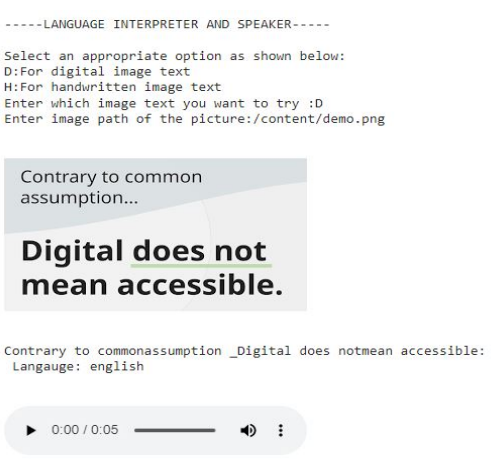

Fig. 2. Digital Image

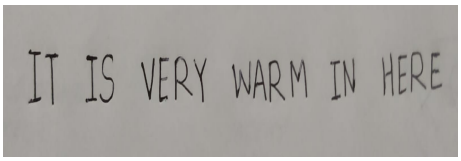

Fig. 3. Handwritten sentence

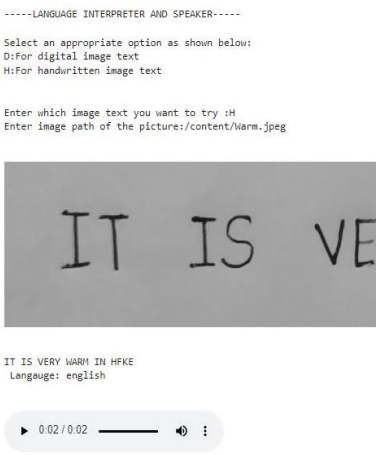

Fig. 4. Handwritten Image

Fig. 5. Invalid option selected

VI. Future Work

Currently, the project only recognises the English language. However, in the future, we will be able to recognise all Indian languages, as well as other major languages, and convert them into any language the user desires. Also, we can add a feature of choosing a language for audio conversion. For visually challenged people we can add voice commands for all processes.

For more precise language recognition, we can add another model for language detection and train it on a larger dataset. The word beam search approach, which predicts the correct word, even if the spelling is erroneous, can be used for more accurate word prediction.This project can be included as a part of many big applications and websites as it will be helpful for visually impaired people in very important aspects and will make them more independent.

VII. CONCLUSION

Several research on text conversion and audio converters have been published recently, however none of them include all of the procedures. As a result, this project demonstrates the effort put forth to extract text from an image using optical character recognition and then identify the text's language. The camera captures the image and is loaded on the system. The OCR engine does additional processing on the captured image and produces the converted text on the screen by using the AlexNet model of CNN algorithm, which gives very good accuracy (i.e 91 percentage on validation data) with good handwriting accuracy varies from handwriting to handwriting. Then from the text language identification is done using some inbuilt libraries such as the stopword library and then, the GTTS library is used to transform the text to audio. This proposed application is both cost-effective and user-friendly, as well as being real-time. It would undoubtedly assist people in determining which language was used to compose the given content, as well as blind persons in reading the documents.

VIII. ACKNOWLEDGEMENT

We would like to convey our heartfelt gratitude to Prof. Sainath Patil, my internal guide, for providing all necessary assistance and support in ensuring the successful implementation of the proposal. We would also like to thank our parents and friends for their assistance in completing the paper in a limited time frame. We are equally grateful to our faculty of management for their support. This paper includes collective efforts and dedication from our group members The success and outcome required a lot of guidance from many people and we are very fortunate to have got all this help.

References

[1] Vaibhav. V. Mainkar, Jyoti A. Katkar, Ajinkya B. Upade, Poonam R. Pednekar. "Handwritten Character Recognition to Obtain Editable Text" , 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020

[2] N. Jayanthi, Harsha Harsha, Naman Jain, Ishpreet Singh Dhingra. "Language Detection of Text Document Image" , 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), 2020

[3] Sneha.C. Madre, S.B. Gundre. "OCR Based Image Text to Speech Conversion Using MATLAB" , 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018

[4] Mayur Bhargab Bora, Dinthisrang Daimary, Khwairakpam Amitab, Debdatta Kandar,Handwritten Character Recognition from Images using CNN-ECOC,Procedia Computer Science,Volume 167,2020, Pages 2403-2409,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.293.

[5] Choudhary, A., Rishi, R., and Ahlawat, S., "Off- Line Handwritten Character Recognition using Features Extracted from Binarization Technique", AASRI Conference on Intelligent Systems and Control, 2013, pp. 306-312.

[6] Handwritten character - https://www.kaggle.com/vaibhao/handwritten-characters

[7] IAM-Dataset - https://www.kaggle.com/datasets/naderabdalghani/iam-handwritten-forms-dataset