Language Interpreter and Speaker

Ruchi Bari[1], *Mrunmayee Apte*[2], *Aakanksha Mohite*[3], *and Sainath Patil*[4]

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 generation, we are heading towards a phase where computers would be capable of doing all 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 once learning is complete, it should be able to recognize the language. The project is divided into three parts. Initially, the handwritten image text would be converted to normal text. In the second part, the language would be identified from the converted text and last, the text would be converted to audio format. This paper discusses the implementation of this idea, gives an approach to problems and challenges that we came across, and some solutions.

Keywords : AlexNet, CNN (Convolution Neural Network), gTTS (google-text-to-speech), Image Processing

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 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. There are around 6900 languages spoken all around us and each is unique in a number of ways. There is no person on the Earth who knows all the existing languages. Thus, knowing the language of the text is the primary step of any type of communication. The present 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. To overcome this problem, this paper gives a solution of transforming the 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

Manuscript Received : February 20, 2022 ; Revised : March 8, 2022 ; Accepted : March 16, 2022. Date of Publication : April 5, 2022.

R. Bari1[1], *Student*, Email : ruche.181914205@vcet.edu.in ; ORCID iD : https://orcid.org/0000-0002-9315-6674
M. Apte[2], *Student*, Email : mrunmayeeapte27@gmail.com ; ORCID iD : https://orcid.org/0000-0002-0770-0400
A. Mohite [3], *Student*, Email : aakanksharmohite@gmail.com ; ORCID iD : https://orcid.org/0000-0002-1077-0100
S. Patil[4], *Assistant Professor (Guide),* Email : sainath.patil@vcet.edu.in ; ORCID iD : https://orcid.org/0000-0002-8226-43895
[1,2,3,4]Vidyavardhini College of Engineering and Technology, K.T. Marg, Vartak College Campus, Vasai Road (W), District Palghar, Vasai, Maharashtra - 401 202.

DOI: https://doi.org/10.17010/ijcs/2022/v7/i2/169682

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 takes lot of time. Also, we cannot use this method on digital images. So, 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, then using the CNN model it 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 the following papers:

In [1] the authors designed a system using Optical Character Recognition (OCR). Pre-processing, segmentation, feature extraction, and post-processing 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% 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 recognized. A choice is made and it must be stored in proper location. When text is printed, it is more accurate than that in handwritten form.

In [2] the authors proposed a technique to detect the language of a text document image that contains Indian languages. 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 13,000 images, each of English, Hindi, and Tamil were taken. The network is trained on characters from various languages, and accuracy of roughly 74% 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 the input text.

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

With the aid of CNN, the developer of [4] detailed handwritten character recognition from photos. 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.

In [5] the researchers first performed Binarization technique on an image which was followed by feature extraction for offline Handwritten Character Recognition. Binarization helps to extract features of handwritten English characters. The algorithm used gave classification accuracy of 85.62%.

For this project, datasets were obtained from Kaggle Handwritten Character [6] and IAM-Dataset [7]. In [2] the researchers created their own dataset after collecting images from different websites. Also, there are different types of Convolutional Neural Networks which are used for OCR. In [4] the researchers performed research on accuracy of different CNN models such as AlexNet, Lenet, Zfnet, etc. 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 follows:

(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 is a very important step of our project as it is 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 made use of CNN - Alexnet model. We made two sections of the dataset: training and validation. There were 1,40,000 images in training and 15,209 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 is obtained in the first part. For language identification, we 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 into speech format. To achieve this, we made use of the gTTS API. By using this, the text is converted into audio format.

The closing step of the project is the UI/UX. For UI, Stream - lit is used. It is nothing but a framework to construct web apps for ML and Data Science. Python is a programming language. It is 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.

A. Handwritten Text Recognition

To convert text from handwritten text image to digital text the following steps were followed:

1) Image Gathering: The system takes an image taken using a mobile phone and executes image processing procedures on it. This system also crops the extra white spaces around the image which helps better the contouring . Fig. 1 shows flowchart of the process.

2) Preprocessing and Segmentation: A handwritten text image is more sensitive to noise, therefore 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

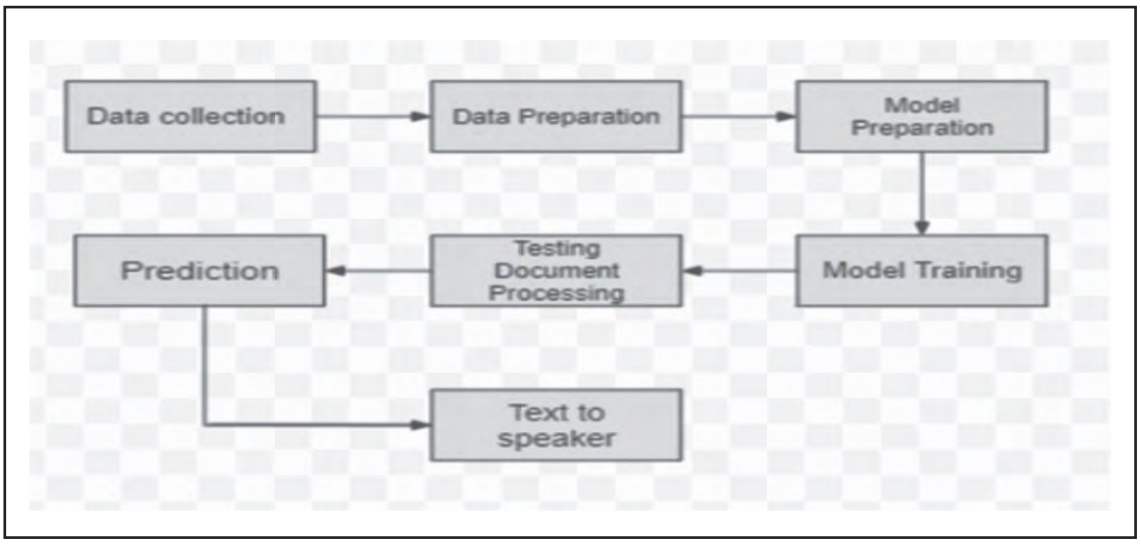

Fig. 1. Flowchart of the Process

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.

3) Feature Extraction: The procedure in which OCR recognizes 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 with 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 recognized 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 recognized.

4) Contours: These 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.

5) Post-Processing: Only the computer understands the extracted output. As a result, data must be saved in a specific format (.txt). The ASCII data was created from the recognized data. The following steps are involved in this work:

(i) Image acquisition using an Android camera.

(ii) Loading the image on the Android Studio-created Graphical User Interface (GUI).

(iii) Preprocessing of the image.

(iv) Feature extraction from the input image.

(v) OCR is used to turn recognised data into text format.

B. 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. We made use of stopwords for doing this .

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. Stopwords 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) 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 given sentence, as the words present in the sentence give us the meaning of the sentence rather than considering the 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 and it can also help us get information about the frequency of a particular word in the sentence.

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 is one of the models of CNN. CNN is a kind 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 architecture similar to that of LeNet, but AlexNet is deeper and has more filters.

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

Table II shows some of the models that we tried on the dataset.

C. *Converting Text Into Speech*

The conversion of text to speech is the process of

TABLE I.

ARCHITECTURE OF ALEXNET

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 II

ACCURACY OF DIFFERENT MODELS USED

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

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 analyed using Natural Language Processing and digital signal processing.

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 that one must have a version of Python greater than 2.7.

V. RESULT AND DECISION

The results of implementing the language of text and audio file of a specified text on a user input image are shown in Fig. 2 to Fig. 5.

VI. FUTURE WORK

At present the project only recognizes the English language. However, in future it will be possible to recognize all Indian languages, as well as other major foreign 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.

Fig. 2. Digital Image

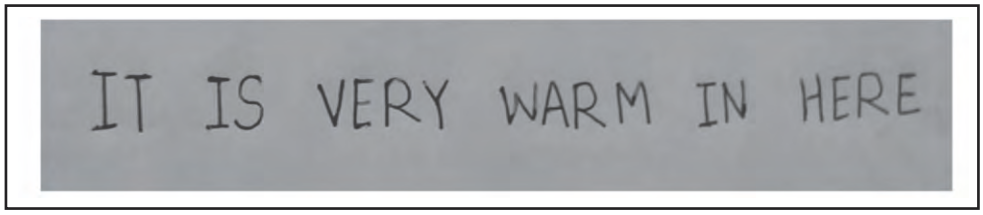

Fig. 3. Handwritten Sentence

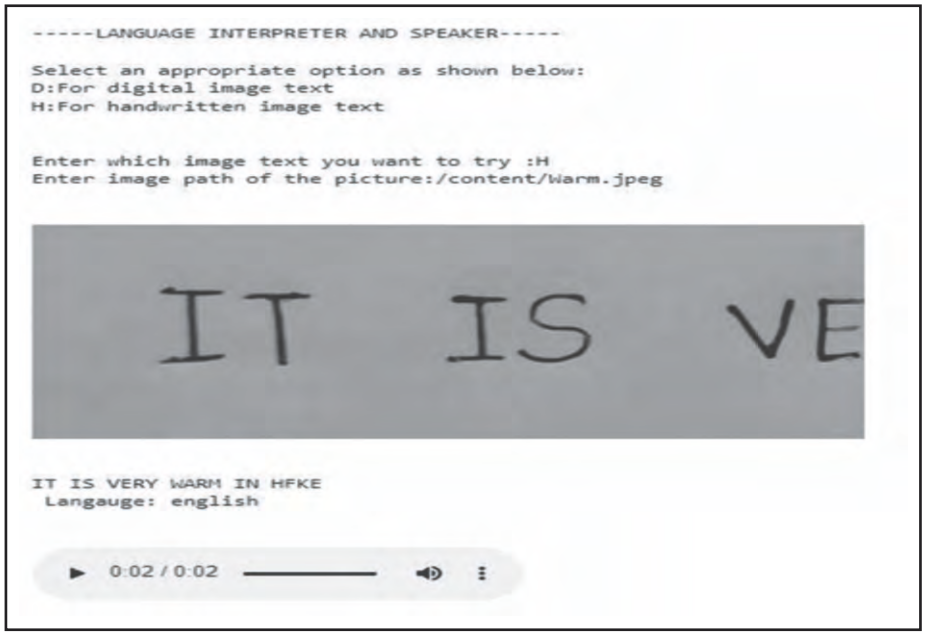

Fig. 4. Handwritten Image

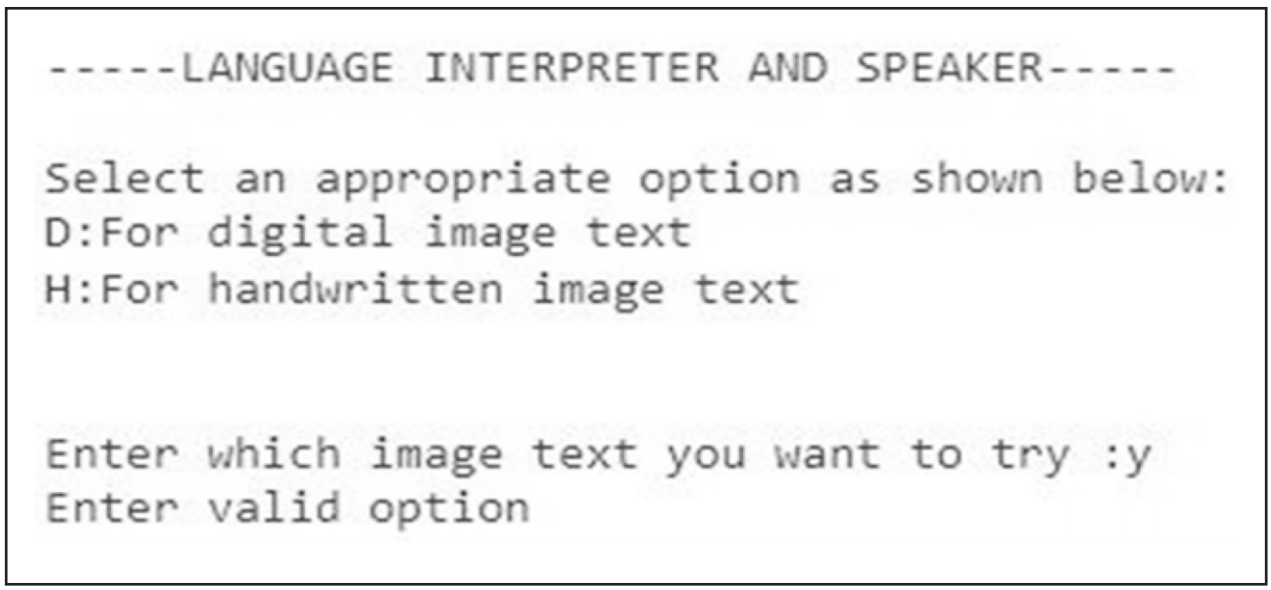

Fig. 5. Invalid Option Selected

VII. CONCLUSION

Several researches on text conversion and audio converters have been published recently, however, none of them includes all of the procedures. This project demonstrates the effort put in to extract text from an image using OCR 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 (91% on validation data) with good handwriting. Accuracy varies from handwriting to handwriting. Then from the text language identification is done using some in-built libraries such as the stopword library and the gTTS library is used to transform the text to audio. This proposed application is both cost-effective and user-friendly, as well as real-time. It would undoubtedly assist people in determining which language was used to compose the given content as well as help blind persons in reading documents.

VIII. ACKNOWLEDGEMENT

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

AUTHORS' CONTRIBUTION

All the authors have performed the work described in this paper. Mrunmayee Apte worked on the different models for handwritten text reorganization and checked the accuracy of the model. Ruchi Bari worked with the language identification and text to speech conversion. Aakanksha Mohite worked on the dataset and developed GUI. Sainath Patil guided this work.

CONFLICT OF INTEREST

The authors have no affiliation with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter, or materials discussed in this manuscript.

FUNDING ACKNOWLEDGEMENT

The authors have not received any financial support for the research, authorship, or for publication of the article.

REFERENCES

[1] V. V. Mainkar, J. A. Katkar, A. B. Upade, and P. R. Pednekar, "Handwritten character recognition to obtain editable text," in *2020 Int. Conf. Electronics Sustainable Communication Syst.*, 2020, pp. 599–602, doi: 10.1109/ICESC48915.2020.9155786

[2] N. Jayanthi, H. Harsha, N. Jain, and I. S. Dhingra, "Language detection of text document image," in *2020 7th Int. Conf. Signal Process. Integr. Networks (SPIN)*, 2020, pp. 647–653, doi: 10.1109/SPIN48934.2020.9071167

[3] S. C. Madre and S. B. Gundre, "OCR based image text to speech conversion using MATLAB," in *2018 2nd Int. Conf. Intelligent Computing Control Systems*, 2018, pp. 858–861, doi: 10.1109/ICCONS.2018.8663023

[4] M. B. Bora, D. Daimary, K. Amitab, and D. Kandar, "Handwritten character recognition from images using CNN-ECOC," *Procedia Comput. Sci.*, vol. 167, 2020, pp. 2403–2409, doi: 10.1016/j.procs.2020.03.293

[5] A. Choudhary, R. Rishi, and S. Ahlawat, "Off-line handwritten character recognition using features extracted from Binarization technique," *AASRI Procedia,* vol. 4, pp. 306–312, 2013, doi: 10.1016/j.aasri.2013.10.045

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

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

About the Authors

Ruchi Bari is pursuing B.E. (Information Technology) from the University of Mumbai. Her research interests include Machine learning, Artificial Intelligence, and Frontend Development.

Mrunmayee Apte is pursuing B. E. (Information Technology from the University of Mumbai. Her research interests include Deep Learning, Artificial Intelligence, and CNN.

Aakanksha Mohite is pursuing B. E. (Information Technology) at the University of Mumbai, Deep learning, artificial intelligence, and CNN are among the topics of her research interest. She is also interested in front-end development.

Sainath Patil has completed B.E. (Electronics) and M.E. (Computers) from University of Mumbai. He is pursuing Ph.D. He is interested in Cyber Security and Machine Learning.