Sign Language to Text and Speech Conversion

A Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Technology
in
Information Technology

by
Abhinav (20168013), Aakarsh Verma (20168002),
Abhishek Dixit (20168004) and Manoj Mahour (20158048)
Group: **IT-21**

to the
Computer Science and Engineering Department
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY
ALLAHABAD, PRAYAGRAJ
November, 2019

UNDERTAKING

I declare that the work presented in this report titled "*Sign Language to Text and Speech Conversion*", submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology Allahabad, Prayagraj., for the award of the ***Bachelor of Technology*** degree in ***Information Technology***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

November, 2019
Allahabad

(Abhinav 20168013)

(Aakarsh Verma 20168002)

(Abhishek Dixit 20168004)

(Manoj Mahour 20158048)

CERTIFICATE

Certified that the work contained in the report titled "*Sign Language to Text and Speech Conversion*", by *Abhinav, Aakarsh Verma, Abhishek Dixit and Manoj Mahour*, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Er. Manoj Wairya)
Computer Science and Engineering Dept.
M.N.N.I.T. Allahabad

November, 2019

Preface

This report is to present the project prepared with the title “Sign language to Text and Speech conversion”. In this report, we will present the need of the project, the technologies used, the results and the expected impact.

During the development of this project, we experimented with many different algorithms and techniques on how to solve the problem at hand, at last settling for one which would be scalable and viable in the future too.

Acknowledgments

After an intensive period of learning and developing, this note of acknowledgement is our final touch on our report. We would like to express our deep gratitude and sincere thanks to everybody who has helped us in completing the project. First and foremost, we would like to thank our mentor and supervisor, Er. Manoj Wairiya, for giving us this opportunity and providing constant support, guidance and encouragement. His innovative ideas and zeal to motivate and help us have led to the successful completion of this project. He has provided us ample opportunity to explore the content and dimensions of this project. We felt privileged working under him.

We would also like to express our sincere gratitude to Prof. Rajeev Tripathi, Director, MNNIT Allahabad, Prayagraj, Prof. Anil Kumar Singh, Head, Computer Science And Engineering Department and Dr. Shashank Srivastava, Head, Departmental Under Graduate Committee (DUGC) for providing us with the tools and facilities to complete the project.

Finally, we would like to thank our friends and family for their constant support and advice. Without their love, blessings and encouragement it would have proved impossible to complete this project.

Abstract

Sign language is the only medium of communication between the people with disabilities in hearing and speaking. Even if they need to present their thoughts or ideas to any individual, they will do that using actions. It may not be understood accurately and efficiently by the individual, which may result in misunderstanding leading to greater problems.

In this project, we have aimed at converting their actions, the sign language, to text. The actions will be understood by the system, depending upon the knowledge of the actions, and then translated accordingly to the required text. The text can be easily read and thus favoring communication.

The software also aims at saving time by encoding some signs to a text that would normally be large in size. We just need to use the desired sign and it will be converted automatically to the large text. This can also be used for security purposes, for developing one's own Sign language.

Contents

Chapter 1

Introduction

Today, there are almost 2 million people classified as Deaf and Dumb. They have great difficulty in communicating with each other and with other individuals as the only means of communication is sign language. They need to learn this sign language. It is extremely difficult for a person who is unaware of this sign language to understand and decode their actions.

It is impossible to identify anything without it's prior knowledge. Even for computers, they need to have information in their memory to identify and provide data related to any object. Now one particular object may differ from another similar type of object in shape, size, orientation or even visual effects may differ. But all the different forms of 1 type of object must be classified in the same category.

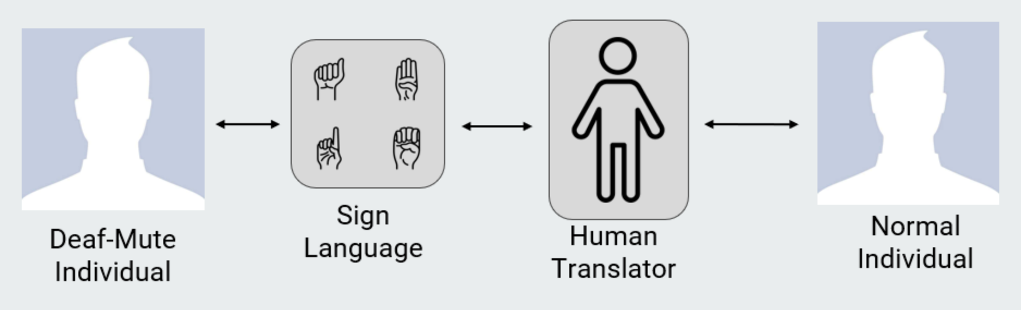


Figure 1: Present Scenario

Making this our aim in abstracting,modifying,analyse and identify the various signals used by the Deaf and Mute to communicate, we have developed this model. The major technologies used are IMAGE PROCESSING and CNN (Convolutional Neural Network).

A number of images of some gestures are taken and processed to make the dataset. The CNN model is then trained using Keras on these captured and modified images. The signs to be translated are then fed to the software which matches it with the existing images and classifies it.

1.1 Motivation

It is very difficult for the deaf people to communicate with the hearing person and there are not many options available to help them. And all of the alternatives have some major flaws.Interpreters are not usually available and are expensive. Pen and paper is also not a good idea, it is uncomfortable, messy and even time consuming, both for the deaf and the hearing person. With the evolution of IoT, everything around is getting automated. The demand for Machine Learning and it's applications is very high. The accuracy and efficiency of any algorithm and the model developed must be very high to make it useful.The knowledge of Machine Learning thus becomes very important.

In the era of Machine Learning where everything is getting automated, the need for an Interpreter to translate Sign language to text is just waste of resources. The classification of objects, object detection and image processing plays a very vital role.

Thus, the main aim is to bridge the gap between the normal and the deaf-and-mute individuals by providing an automatic translation system.

Chapter 2

Related Work

Text classification is one of the most used application of Machine Learning. It is used to automatically assign predefined categories(labels) to text documents. The purpose of text classification is to organize conceptually a large collection of documents. It has become more relevant with exponential growth of the data, with wide applicability in real world applications.

There are apps available in the market for converting signs into text, but all of them use old technology are slow to operate. Messengers and texting are used, but the problem is still not solved, which is translation,and do not offer neat, confident and comfortable way to communicate.

Google has developed an app called GnoSys [2], that uses neural network and computer vision to recognise the video of sign language speaker,and then smart algorithms translate it into speech.

A glove was developed at the University of California, San Diego,in July 2017 which can convert the 26 letters of American Sign Language (ASL) into text on a smartphone or computer screen.But it was limited only to 26 letters of English alphabet.

Start-ups working with NAD (National Deaf Association) have collaborated with India Accelerator to gather sign language data for India. All the available apps have limited vocabulary of signs and hence research is going on in this field.

Chapter 3

Proposed Work

3.1 Objective

Deaf people do not have that many options for communicating with a hearing person, and all of the alternatives have some major flaws. Interpreters are not available easily, and also can be expensive. Affordable and always available interpreter services are in huge demand in the deaf community. Every day thousands of local businesses around the globe face problems with providing their services to deaf.

According to the National Deaf Association (NAD), 18 million people are estimated to be deaf in India.[2]

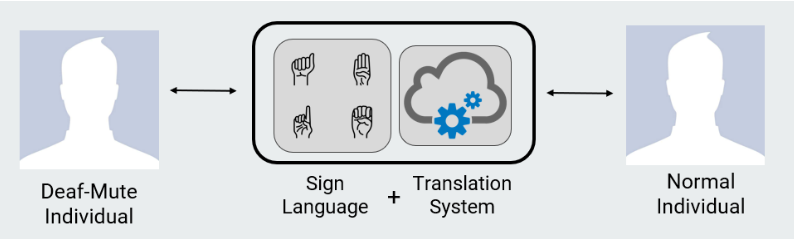


Figure 2: Objective of the Solution

The main objective is to design a software that will help drive inclusivity at the workplace by removing communication barriers between the disabled and able.The new application can find use in a B2B setting, where businesses who want to employ deaf and mute employees can use it to convey employee messages to the end consumer.

An easy to use innovative digital translator that is compellingly fast, easy, comfortable and economical is the need of the hour.

3.2 Software Requirement Specification

3.2.1 Introduction

▮ *Purpose*

The aim of this document is to provide a detailed description of the translator of Sign language to text. It will cover the applications and features of the system, the interfaces of the system, what the system is expected to do, the constraints that the project will work under and how it behaves in response to external stimuli. This is intended for both the developers and the users of this system.

▮ *Scope*

This system is primarily intended for making a Interpreter. This will have applications in Business who want to employ deaf and mute employees can use it to convey employee messages to the end consumer. It will be used majorly by the deaf and mute to comunicate.

The applications can further be extended to security purposes, by developing a sign language of your own. And even observing and analysing any suspicious actions.

▮ *Glossary*

- Feature: Features are individual measurable property or characteristic of a phenomenon being observed. These require classification.
- Label: Labels are the final output. We can also consider the output classes to be the labels.
- Model: A machine learning model is a mathematical portrayal of a real-life problem. There are various algorithms that perform different tasks with different levels of accuracy.

- Regression: Regression is a statistical method that is used to predict real and continuous valued functions.
- Classification:In classification, we will need to categorize data into a finite number of predefined classes.
- CNN : It is a Machine Learning unit algorithm, for supervised learning, which is used in classificaton of large amount of data[1].
- Image Processing : The various modifications done on a raw image to make it suitable for the training model.
- Trainig-set : This is the data set over which CNN model is trained. The predictions are completely dependent on the training-data set.

Overview of Document

The next section, the Overall Description section, of this document gives an overview of the functionality of the project. It describes the informal requirements and is used to establish a context for the technical requirements specification in the next section. The third section, Requirements Specification section, of this document is written primarily for the developers and describes in technical terms the details of the functionality of the product. Both sections of the document describe the same software product in its entirety, but are intended for different audiences and thus use different language.

3.2.2 Overall Description

- *System Architecture*

The main objective of the software is to classify the gestures and label them with one of the various categories already defined while training the dataset. We have then tested the same for some data with the help of deep learning.

The system has been trained on previously captured gestures which have been labeled with the text associated with them. Multiple copies of each gesture image have been created to extract the features efficiently, then the CNN model is trained. The new gesture which has to be translated, has to be signaled in front of the camera connected to the system, which will be recorded and matched and classified using CNN algorithm.

3.2.3 Requirement Specification

- *Functional Requirements*

The Prime objective of the software is to translate the sign language into text. Initially, one character was translated at a time and later, the software was trained and developed to translate even words. Words can be formed by concatenating various charachters as well and the formed word will be displayed on the output window.

The captured images need to be pre-processed. The system modified the images captured and trained the model to classify the signals in one of these defined labels.

- *Non-Functional Requirements*

The sole purpose of the software is to facilitate communication for the disabled. The previous available devices were slow and inefficient. Thus, the software is built to translate the signs accurately and at a relatively faster rate. The software is designed efficiently such that it can be modified easily making it easy to maintain.

3.3 Overview of Approach

We have created an User Interface which allows the User to add new gestures with some specific meaning, and the option to train the model with the added features. The UI also directs the user to "hand recognition" window, which will further be used to do gestures. Then finally two options are provided for translating sign language to Text, one allowing one character at a time, and second to concatenate the characters to form a complete word.

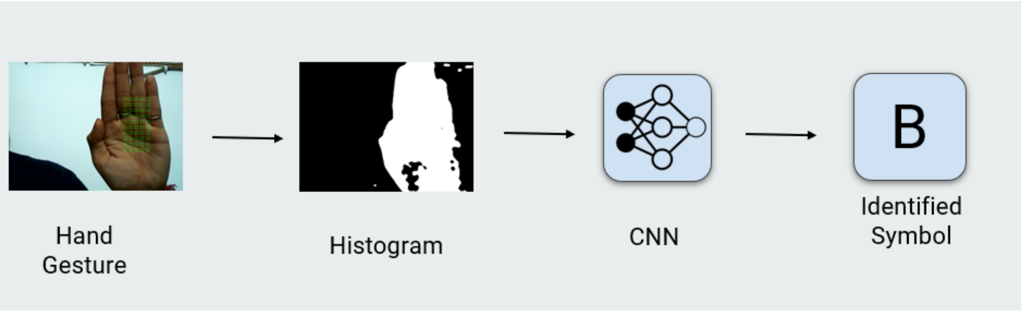


Figure 3: Flowchart of the solution

There are a total of 44 gestures for which the model is trained, including 26 English alphabets, 10 numberic digits and 8 other commonly used symbols. 1200 images of each gesture have been captured and then each image is flipped, making a total of 2400 images for each gesture. The images have then been resized to 50*50 pixels and converted to grayscale.

A histogram has been created which identifies the skin of the hand of the speaker and separates it from the background[3]. The Neural Network is trained using Keras on each gesture. Whenever the translator is to be used, the person uses the sign language to be recorded in the camera of the device with the software, which then fetches it into the Neural Network and the sign is classified.

3.4 Pre-processing the Data

3.4.1 Histogram creation

The OpenCV library has been used to generate a histogram that will separate the hand gestures from the background. For this purpose 50 squares in the form of 5*10 are displayed and the hand must cover all the squares. Then the image is captured and a histogram is plotted of the area covering the squares.

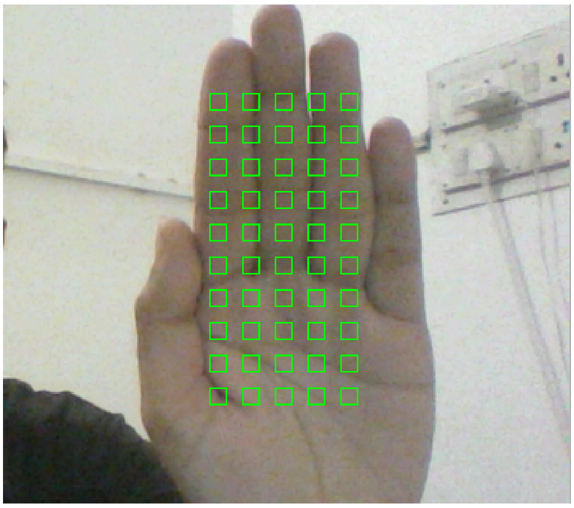


Figure 4: Hand detection

3.4.2 Image Processing

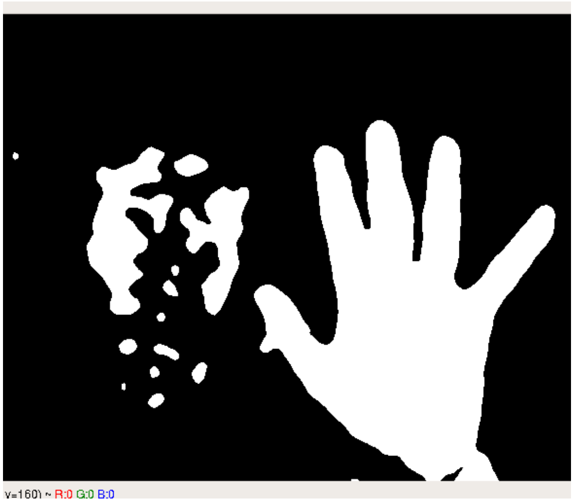


Figure 5: Histogram of the hand

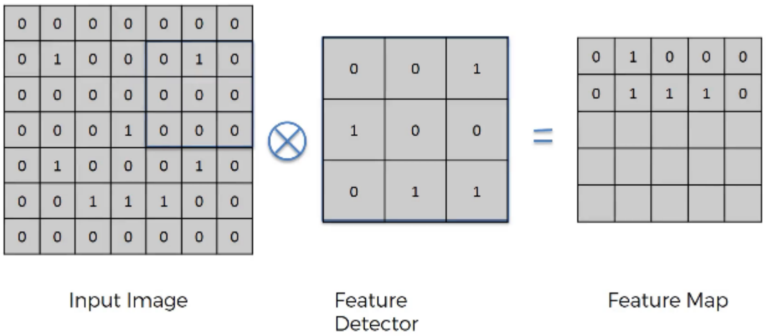
The images captured of the gestures using the intensities obtained from the histogram are then processed. The images are resized to 50*50 pixels and then converted to gray scale.Each image is then flipped along the vertical axis.

3.4.3 Convolution

The image obtained is then convoluted with the feature detector to form the feature map[1]. This is the most important part in feature detection.The image is convoluted with a number of features and hence a number of feature maps are present.Larger the number of features, better it is to classify the image.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \qquad (1)$$

The main aim of convolution is to detect features.



0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

0	0	1
1	0	0
0	1	1

0	1	0	0	0
0	1	1	1	0

Figure 6: Convolution of a matrix

3.4.4 ReLu Layer

The Convoluted image is then passed through the Activation function. The activation function used is the RECTIFIER FUNCTION.This is used to increase the non-linearity in the image.

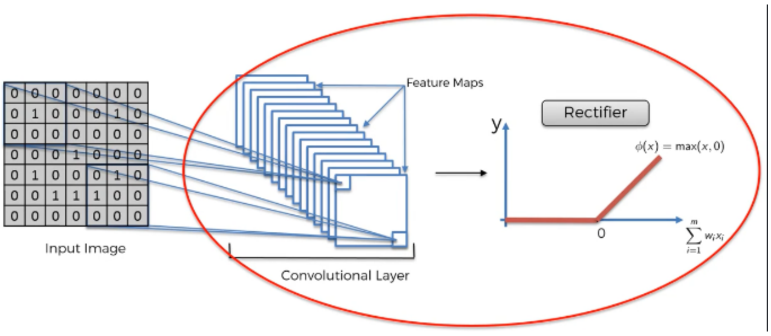
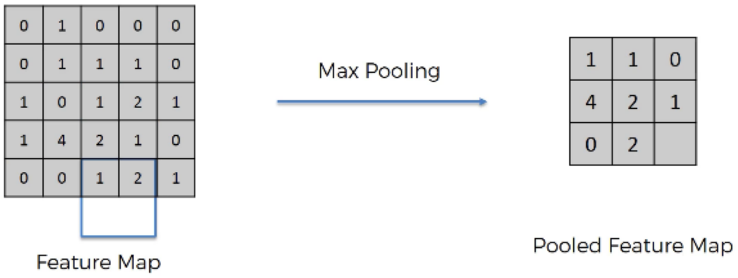


Figure 7: Activation function: Rectifier function

3.4.5 Max pooling

Now the feature map contains the convoluted result. There are a large number of such maps, hence enormous data. Further, one feature may differ in size, orientation in different images. Both these issues are resolved using Max Pooling. A small grid is selected and then the maximum value is preserved, reducing the size of data as well.



0	1	0	0	0
0	1	1	1	0
1	0	1	2	1
1	4	2	1	0
0	0	1	2	1

1	1	0
4	2	1
0	2	

Figure 8: Max Pooling

3.4.6 Flattening

The pooled image features need to be flattened so that they can be used as input in the next Artificial Neural Network. These form the input layer of the Artificial Neural Network.The max pooling output is transformed into a column.

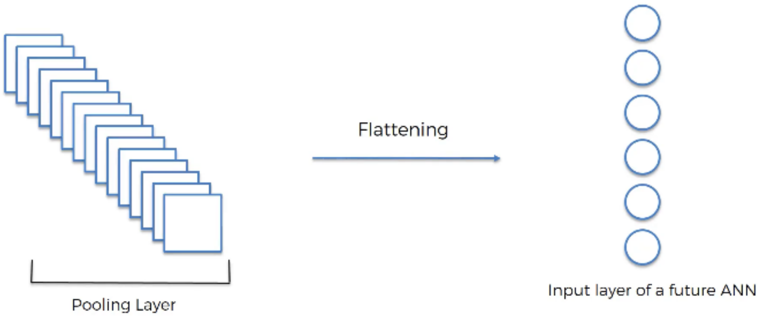


Figure 9: Flattening

3.5 Building the Model

The CNN is a type of Artificial Neural Network that is used in Image recognition and processing. It detects specific features in the images and then classifies them accordingly based on the presence of those features.In our model we have 5 layers in total including 3 Convolutional layers and 2 fully connected Dense layers.

The first CNN layer consists of 16 neurons, the second layer consists of 32 neurons and the third layer consists of 64 neurons.

The first dense layers consists of 128 neurons and uses the Rectifier function as the Activation function. The second dense layer uses Softmax function as the Activation function.

3.5.1 Activation functions

Activation Function is the function that is applied to the sum of weighted inputs to the neuron. This is where calculations happen. In our neural network, we have used two different activation functions.

- Relu Activaton Function: ReLU stands for Rectified Linear Unit. Its cheap to compute as there is no complicated math[1]. The model can therefore take less time to train or run. This is used to increase the non-linearity of the images.It is especially useful when dealing with small values as in our case.The formula for ReLu function is given by:

$$y = max(0, x) \tag{2}$$

- Softmax Activation function:Softmax is an activation function. It is frequently used in classifications. Softmax output is large if the score is large. Its output is small if the score is small. The proportion is not uniform. Softmax is exponential and enlarges differences. The formula for Softmax function is give by:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3}$$

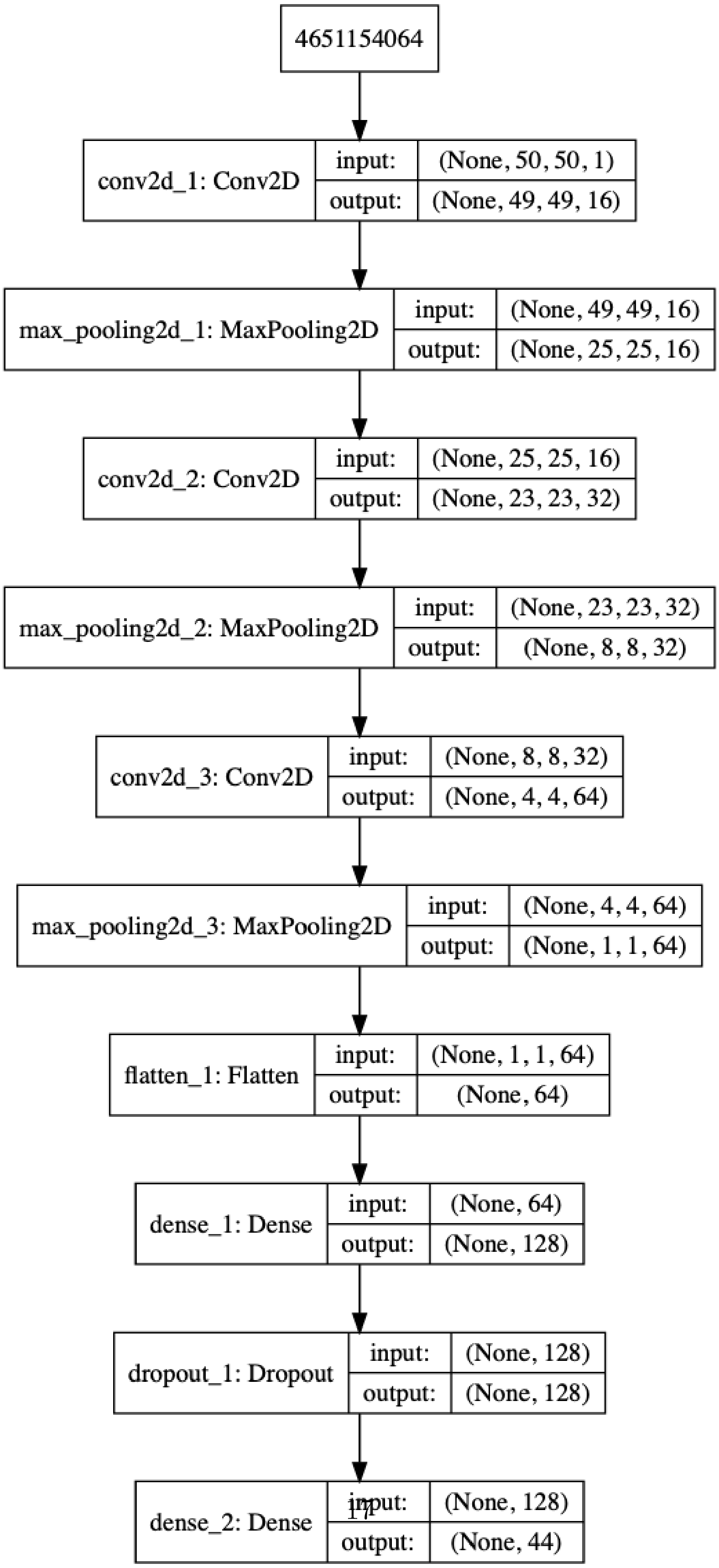


Figure 10: Model consisting of 5 layers

3.6 Data Flow Diagram

3.6.1 Zero level DFD

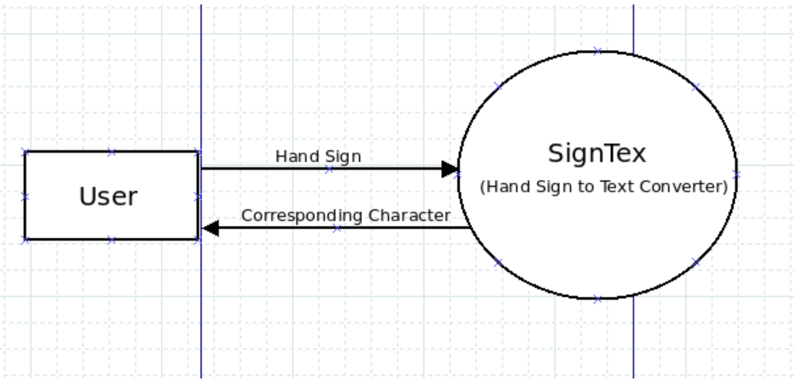


Figure 11: Zero Level DFD

3.6.2 First level DFD

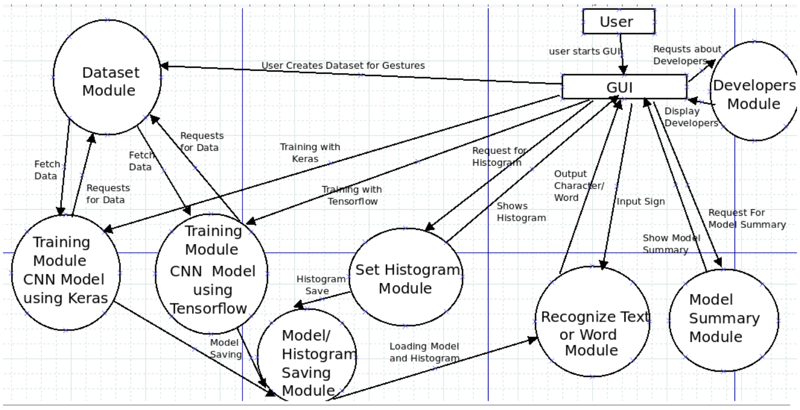


Figure 12: First Level DFD

3.7 User Interface

An user friendly User Interface is created using Tkinter. It provides buttons for various operations including Adding more gestures, Training the model, Setting the histogram, Open the translator in two modes, and provides an option to view the model summary as well.

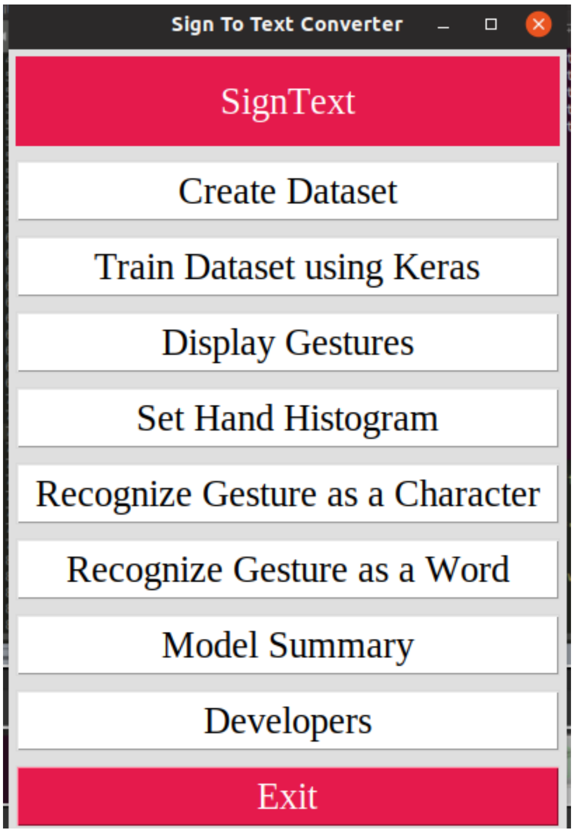


Figure 13: User Interface

Chapter 4

Results And Analysis

4.1 Results

The model is translating the sign language to the English alphabet characters precisely. Some of the translated outputs are:

4.1.1 Recognizing gesture as character

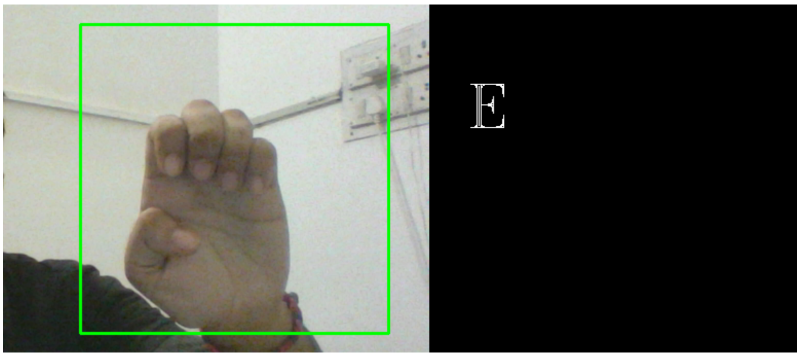
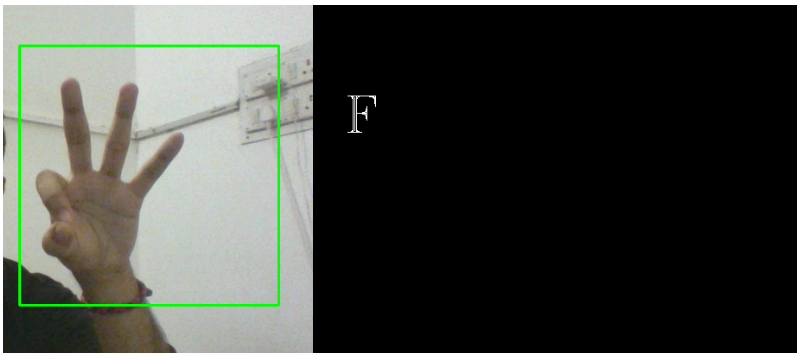


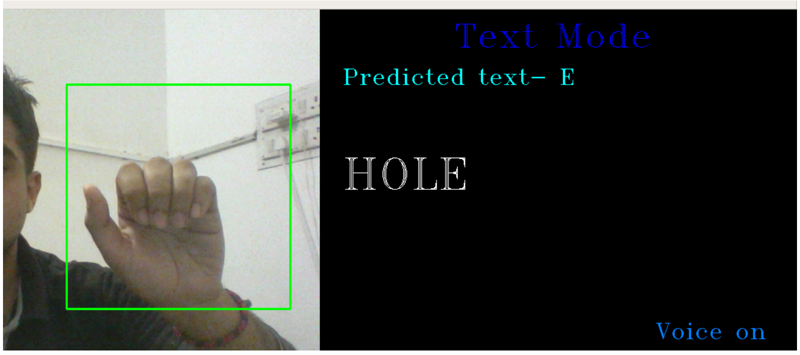
Figure 14: Sign Language to Text for 'E'



F

Figure 15: Sign Language to Text for 'F'

4.1.2 Recognizing gesture as word



Text Mode

Predicted text- E

HOLE

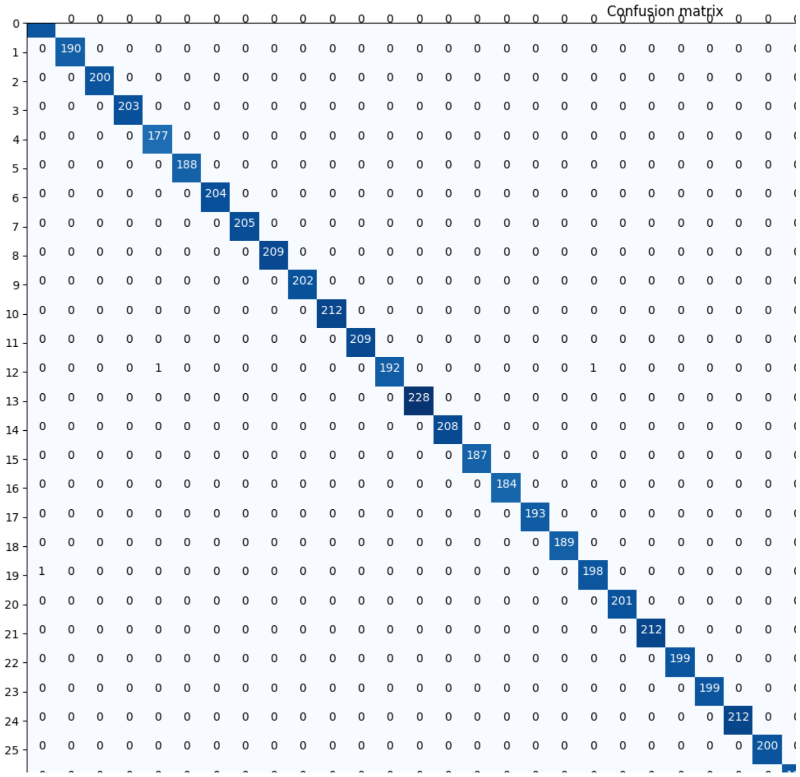
Voice on

Figure 16: Sign Language to Text for 'HOLE'

In this, each character is translated first, then if it stays of sometime, it gets concatenated to the word being formed. The word is made by concatenating one character at a time.

4.1.3 Confusion Matrix

Classification was using Convolutional Neural Network. The accuracy obtained on the test data was 0.9997 and misclass was 0.0003.



0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	190	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	203	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	177	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	188	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	204	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	205	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	202	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	212	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	192	0	0	0	0	0	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	228	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	208	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	187	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	184	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	193	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	189	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	198	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	201	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	212	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	199	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	199	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	212	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200

Figure 17: Confusion Matrix

4.2 Analysis

The accuracy of the translator is increasing as the number of epochs are increased while training the model and when the number of images for each gesture are increased.

Epochs : It is the number of times the model is trained over the same data-set.

4.2.1 Recall

It is the fraction of the total number of relevant instances that were retrieved.

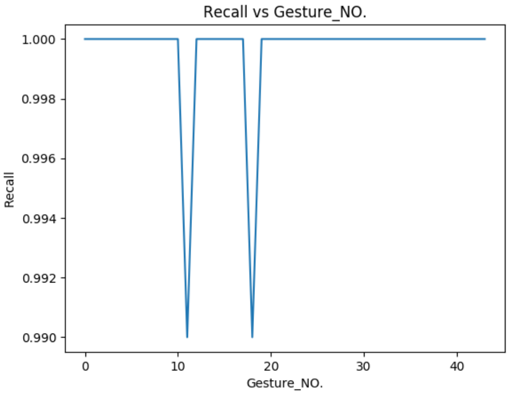


Figure 18: Recall vs Gesture graph

4.2.2 Precision

It is defined as the fraction of the relevant instances among the instances retrieved.

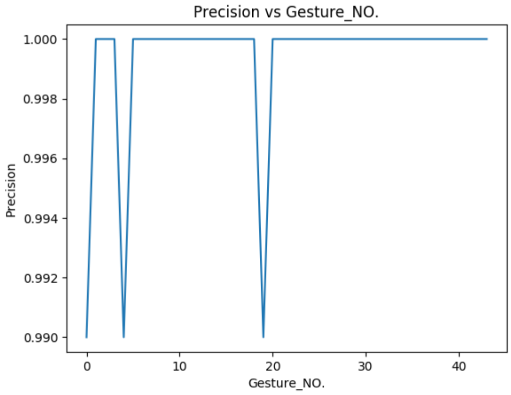


Figure 19: Precision vs Gesture graph

4.2.3 F-1 score

It is defined as the harmonic mean of the Recall and the Precision.In this, even the false negative and false positive are crucial. It tells the accuracy of the classifier in classifying the data points in that particular class compared to all other classes.

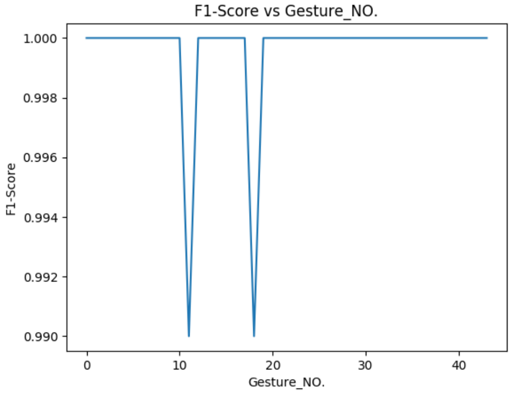


Figure 20: F1 score vs Gesture graph

Chapter 5

Conclusion and Future Work

5.1 Conclusion

A translator that can translate 44 gestures to Text and even concatenate the translated characters to form words was successfully developed. We have presented our results obtained and have plotted the graphs for the precision, recall and F1 scores.The translator can be made more effective by adding more gestures. As suggested and guided by our project mentor, a lot of measures were taken and modifications made to tackle with all the problems coming up in the project that required our undivided attention.

5.2 Applications

1. Deaf and Dumb people will be able to convey their thoughts with hand gestures without the need of a translator.

2. Can be used to detect any kind of threat where sign language is being used to communicate, by analysing the suspected gestures.

3. Sensor based glove can work even in dark and areas with high noise images.

5.3 Limitations

1. System will fail to work in accelerated motions.

2. Image processing based implementation would fail in dark and when background is very similar to skin color.

3. Complicated gestures involving motion cannot be identified.

5.4 Future Scope

1. Adjustment to count for system in motion.

2. Identify complicated gestures involving two hands and motion.

3. Use hand gestures to control and automate other devices.

References

[1] Gradient based learning. http://yann.lecun.com/exdb/publis/pdf, 1998.

[2] Meet the new google translator. https://economictimes.indiatimes.com, 2018.

[3] Sign language to text. http://github.com/EvilPort2, 2018.