SIGN LANGUAGE RECOGNITION AND CONVERSION TO SPEECH

A Project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR'S OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING and

ELECTRONICS AND COMMUNICATIONS ENGINEERING

by

K.Karthik Kumar(111915061)

Mohammed Thousif Ahmed(111916029)

K.Rohit Reddy(111915059)

B.Karthik Reddy(111916063)

Semester: V

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING and ELECTRONICS AND COMMUNICATIONS ENGINEERING

Indian Institute of Information And Technology Pune

Near Bopdev Ghat, Kondhwa Annexe, Yewalewadi, Pune, Maharashtra 411048

NOVEMBER 2021

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **" Sign Language Recognition and conversion to speech "** submitted by **K.Karthik kumar,Mohammed Thousif Ahmed, K.Rohit Reddy,B.Karthik Reddy** bearing the **MIS No: 111915061, 111916029, 111915059, 111916063** respectively, in completion of our project work under the guidance of **Dr.Tanmoy Hazra** is accepted for the project report submission in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science Engineering and Electronics and Communications Engineering in the Department of Computer Science and Engineering, Indian Institute of Information Technology,Pune, during the academic year 2021-22.

Dr.Tanmoy Hazra	**Dr.Tanmoy Hazra**
Project Guide	Head of the Department
Assistant Professor	Assistant Professor
Department of CSE	Department of CSE
IIIT Pune	IIIT Pune

Project Viva-voce held on 12/11/2021

Internal Examiner	**External Examiner**

ACKNOWLEDGEMENT

This project would not have been possible without the help and cooperation of many. I would like to thank the people who helped me directly and indirectly in the completion of this project work.

First and foremost, I would like to express my gratitude to our beloved director, **Dr.Anupam Shukla**, for providing his kind support in various aspects.

I would like to express my gratitude to my project guide **Dr.Tanmoy Hazra**, Assistant Professor, Department of CSE, for providing excellent guidance, encouragement, inspiration, constant and timely support throughout this B.Tech project.

I would like to express my gratitude to the head of department **Dr.Tanmoy Hazra**, Assistant Professor, Department of CSE, for providing his kind support in various aspects.

I would also like to thank all the faculty members in the Dept. of CSE and my classmates for their steadfast and strong support and engagement with this project.

Abstract

Inability to speak is considered to be a true disability. People with this disability use different modes to communicate with others, there are a number of methods available for their communication, one such common method of communication is sign language. Developing sign language applications for deaf people can be very important, as they will be able to communicate easily with even those who dont understand sign language. According to the NAD, 18 million people are estimated to be deaf in India. They can communicate among themselves using sign language but having a conversation with others is a difficult task. Extensive work has been done on the American Sign Language(ASL) and by working on this project we want to translate the sign language into text and sound and enable these people to have an improved conversational experience. This work aims at taking the basic step in bridging the communication gap between normal people and deaf and dumb people using sign language. The image dataset consists of 32 ASL gestures.. The accuracy of the model obtained using the Convolution Neural Network was 99.64percent.

Keywords : ASL, Sign Language Character Recognition, Convolution Neural Network, Computer Vision, Machine Learning

TABLE OF CONTENTS

List of Figures

List of Tables

Listings

Chapter 1

Introduction

Sadly, in the fast changing society we live in, people with hearing impairment are usually forgotten and left out. They have to struggle to bring up their ideas, voice out their opinions and express themselves to people who are different to them. Sign language, although being a medium of communication to deaf people, still have no meaning when conveyed to a non-sign language user.

So, we are putting forward a sign language recognition system. It will be an ultimate tool for people with hearing disability to communicate their thoughts as well as a very good interpretation for non sign language user to understand what the latter is saying.

Identification of sign gestures is mainly performed by the following methods:

- The Glove-based method is in which the signer has to wear a hardware glove, while the hand movements are getting captured.
- Vision-based method, further classified into static and dynamic recognition. Statics deals with the detection of static gestures(2d-images) while dynamic is a real-time live capture of the gestures. This involves the use of the camera for capturing movements.

The Glove-based method, seems a bit uncomfortable for practical use, despite having an accuracy of over 90percent.

1.1 Deep Learning

To know about Deep learning we need to first understand about artificial intelligence. Artificial intelligence (AI) is a very vast branch in Computer Science and it is used for constructing smart machines. The smart machines are capable of executing a task that requires human intelligence. Deep learning and machine learning are learning techniques in Artificial Intelligence. We have achieved many advancements in machine learning and deep learning. Now , as we are discussing Deep Learning. Deep learning is an AI function that works almost similar to the human brain in terms of performing tasks such as processing data and creating patterns and in the terms of decision making. Deep learning is a part of machine learning in artificial intelligence which consists of networks such as deep neural networks which are capable of learning unsupervised data.Consider a human brain: it consists of millions of neurons. A deep neural network (DNN) has several layers in between input and output layers. Each layer consists of several nodes and are similar to neurons of the human brain. To solve a task the system has to process layers of data between input and output. Stimulus must be passed at the input layer in order to execute a task.

1.2 Importance of Deep Learning

- Deep learning has a very significant importance in terms of handling big data.
- Deep learning includes a large number of features and its corresponding process.
- However , the access of the vast amount of data is required for deep learning algorithms to be effective.
- But deep learning models will be overfitted if data is too simple or incomplete.
- Deep learning is very useful for real life applications because of its algorithm's potential at learning.

1.3 Thesis Outline

The thesis is organized as follows:

Chapter 1 provides a general introduction to the thesis.

Chapter 2 says the necessity of this type of projects in the present world.

Chapter 3 talks about various technologies used in this project.

Chapter 4 gives the design and analysis of the entire control system.

Chapter 5 provides the experimental results of the proposed scheme and it analyses the activity of this control system in different conditions.

Chapter 6 concludes our work and gives the future work which can be done to improve this scheme.

Chapter 7 gives the references used to build this project.

Chapter 2

Objectives

Introduction of chapter 2.

2.1 Objectives

The aim of Sign Language Recognition is to automatically describe the signs of people who are suffering with hearing and speaking problems who communicate through Sign language.

To recognize hand gestures from a live video sequence.To make a Real Time prediction for sign language.

To create the application for helping the blind people.

Chapter 3

Literature Review

Literature review of our proposed system shows that there have been many explorations done to tackle the sign recognition in videos and images using several methods and algorithms.

1. Siming He[4] proposed a system having a dataset of 40 common words and 10,000 sign language images. To locate the hand regions in the video frame, Faster R-CNN with an embedded RPN module is used. It improves performance in terms of accuracy. Detection and template classification can be done at a higher speed as compared to single stage target detection algorithm such as YOLO. The detection accuracy of Faster R-CNN in the paper increases from 89.0percent to 91.7percent as compared to Fast-RCNN. A 3D CNN is used for feature extraction and a sign-language recognition framework consisting of long and short time memory (LSTM) coding and decoding network are built for the language image sequences.On the problem of RGB sign language image or video recognition in practical problems , the paper merges the hand locating network, 3D CNN feature extraction network and LSTM encoding and decoding to construct the algorithm for extraction. This paper has achieved a recognition of 99percent in common vo- cabulary dataset.

2. K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4896-4899, doi: 10.1109/BigData.2018.8622141.It Consists of 31 American sign gestures.Inception(CNN) is used for Feature extraction.RNN is used to train on temporal features.

3. Hand Gesture Recognition Using PCA by Mandeep Kaur Ahuja et al: In this paper authors presented a scheme using a database-driven hand gesture recognition based upon skin color model approach and thresholding approach along with an effective template matching with can be effectively used for human robotics applications and similar other applications. Initially, the hand region is segmented by applying the skin color model in the YCbCr color space. In the next stage, thresholding is applied to separate foreground and background. Finally, the template-based matching technique is developed using Principal Component Analysis (PCA) for recognition.

4. In this paper, Pigou used CLAP14 as his dataset [9]. It consists of 20 Italian sign gestures.After preprocessing the images , he used a Convolutional Neural network model having 6 layers for training.He has used Rectified linear Units (ReLU) as activation functions. Feature extraction is performed by the CNN while classification uses ANN or fully connected layer. His work has achieved an accuracy of 91.70percent with an error rate of 8.30percent.

5. In this paper by Herath, H.C.M. , W.A.L.V.Kumari and Senevirathne, W.A.P.B and Dissanayake, Maheshi , a low cost approach has been used for image processing . The capture of images was done with a green background so that during processing, the green colour can be easily subtracted from the RGB colourspace and the image gets converted to black and white. The sign gestures were in Sinhala language. The method that thy have proposed in the study is to map the signs using centroid method. It can map the input gesture with a database irrespective of the hands size and position. The prototype has correctly recognised 92percent of the sign gestures.

6. Let's approach the research done by Rekha, J[5]. which made use of YCbCr skin model to detect and fragment the skin region of the hand gestures. Using Principal Curvature based Region Detector, the image features are extracted and classified with Multi class SVM, DTW and non-linear KNN. A dataset of 23 Indian Sign Language static alphabet signs were used for training and 25 videos for testing.

The experimental result obtained were 94.4percent for static and 86.4percent for dynamic.

3.1 Table

S.NO	AUTHOR	WORK DESCRIPTION	PUBLISH YEAR
1	Siming He	Faster R-CNN with an embedded RPN module is used to improve accuracy, A 3D CNN is used for feature extraction.	2019
2	K.Bantupalli and Y.Xie	American Sign Language using Deep Learning and Computer Vision • Consists of 31 American sign gestures • Inception(CNN) is used for Feature extraction. • RNN is used to train on temporal features.	2018
3	Herath et al	Image based Sign Language Recognition System for Sinhala Sign language. • Low cost approach - using green screen in the background for preprocessing as subtraction would be easy. • Mapping signs using centroid method.	2013
4	Pigou	• Data set used CLAP 14. • Consists of 20 Italian sign gestures. • After preprocessing the CNN model having 6 layers is used for training. • CNN is used for Feature Extraction while classification used ANN • Achieved accuracy - 91.70% , Error rate - 8.30%	2015

Chapter 4

Methodology

The below flowchart shows the system architecture of our project figure reference fig: 4.1 in the page : 8.

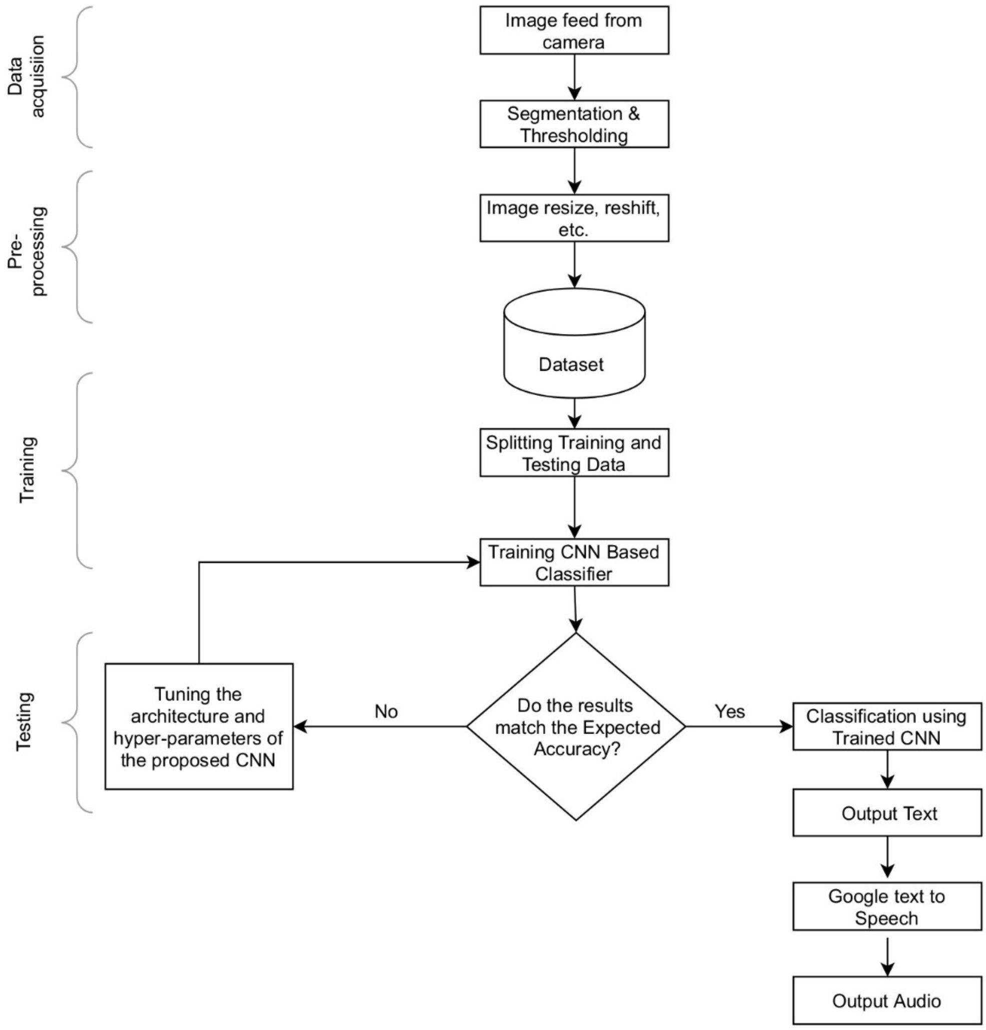

Figure 4.1: Flowchart

4.1 Data Acquisition

We have created our own Dataset by following the American Sign Language for alphabets, special characters and also added few custom gestures like YOLO , All the best, etc.. which can be further extended to add Indian Sign Language Vocabulary. We first segment the Hand region by using Background Subtraction Technique which uses the concept of running averages from a sequence of video frames.Then we do thresholding for this image. By this we can get the contour of the hand region.

Figure 4.2: ASL Alphabet

4.2 Data Preprocessing

For each gesture we get a threshold image and all these images are pre-processed using Keras preprocessing model which does the necessary resizing, shifts, flips, orientations to certain degree, etc.. which concludes our Final Dataset. We split the datasets into two parts. 70percent for training data and 30percent for testing data.

Figure 4.3: Preprocessing

4.3 Feature Extraction

One of the most crucial part in image processing is to select and extract important features from an image. Images when captured and stored as a dataset usually take up a whole lot of space as they are comprised of a huge amount of data. Feature extraction helps us solve this problem by reducing the data after having extracted the important features automatically.It also contributes in maintaining the accuracy of the classifier and simplifies its complexity. In our case, the features found to be crucial are the binary pixels of the images. Scaling the images to 64 pixels has led us to get sufficient features to effectively classify the American Sign Language gestures . In total, we have 4096 number of features, obtained after multiplying 64 by 64 pixels.We are using CNN and PCA for Feature Extraction part.

4.3.1 Convolution Neural Networks

As the name says this model uses Convolution in at least one of it's layers. Here in the first part, we use convolution operation for feature extraction. In this feature extraction we extract important points or features from the image known as Filters.

Pooling: Pooling layers reduce the dimensions of the data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. It also helps with the position invariant feature detection which means in this example no matter where eyes, ears, nose etc. are present, it will detect them. It also makes the model tolerant towards small distortions.

ReLU: It speeds the training of the model.

Padding: padding means giving additional pixels at the boundary of the data.

Flattening: In this process we convert the data into one-dimensional array to give it to the final classification model.

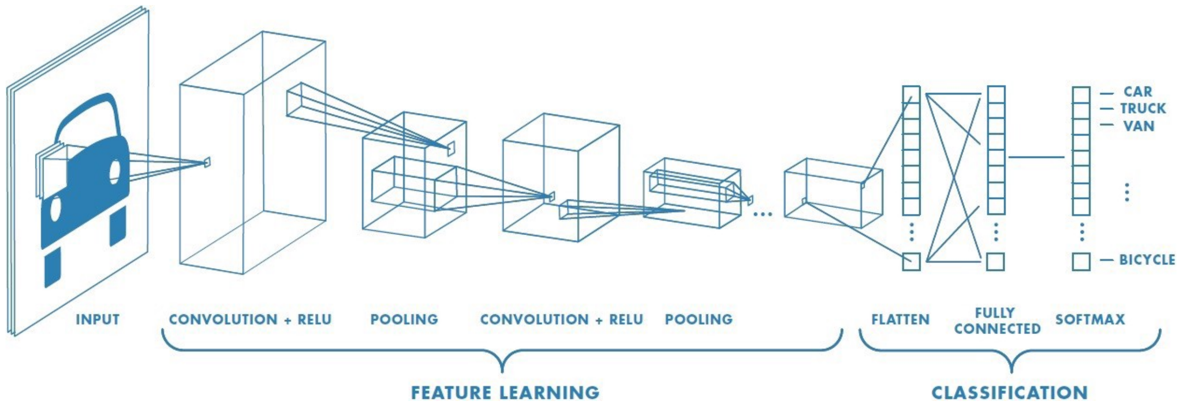

Figure 4.4: CNN architecture

4.3.2 Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality reduction technique that is often used to decrease the dimensionality of large data sets, by converting a large set of variables into a smaller set of variables without losing much information. After performing the preprocessing, Training of the model is performed using PCA. This is mainly achieved by calculating the Covariance matrix and extracting the eigenvector and eigenvalues from the Covariance matrix. The reduced transformation matrix is then calculated while projecting to the PCA space. The mean of all the images in the training dataset is computed to get the reduced transformation matrix. The mean image is subtracted from all the images in the training dataset for calculating the covariance matrix. The covariance matrix is computed from the training dataset, which is used to calculate the eigenvectors and eigenvalues. The eigenvectors and eigenvalues of the dataset images are arranged according to the index in decreasing order. The reduced transformation matrix is calculated and is then used to project the training dataset into the PCA space.

4.4 Model Architecture

4.4.1 CNN

The model architecture includes the Deep Convolutional Neural Network. The network contains 7 hidden convolution layers with Relu as the activation function and 1 Fully connected layer. The network is trained across 50 iterations with a batch size of 64. We kind of saw that 50 iterations kind of trains the model well and there is no increase in validation accuracy along the lines so that should be enough. A convolutional neural network (CNN) is a type of artificial neural network usually designed to extract features from data and to classify given high dimensional data. Pre-processed images are given as input to the CNN model for training. Cross validation method is used to arrive at the best model. The network consists of an input layer, followed by 7 convolutional and max pooling layers alternately and followed by a soft max fully connected output layer to extract features. After extracting features, 2 layer hidden neural-network is used for classification. Dropouts can help to avoid overfitting.The model achieves an accuracy of 99.64percent on the validation dataset and the ratio of training set to validation set is around $210 : 90$.

4.4.2 Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. If the given data is linear we can classify it simply by using line as a hyperplane, but if the data is non-linear we need to add suitable features i.e, add a new dimension. In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyper-plane. No, the SVM algorithm has a technique called the kernel trick. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts non separable problem to separable problem. It is mostly useful in non-linear separation problems. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined.

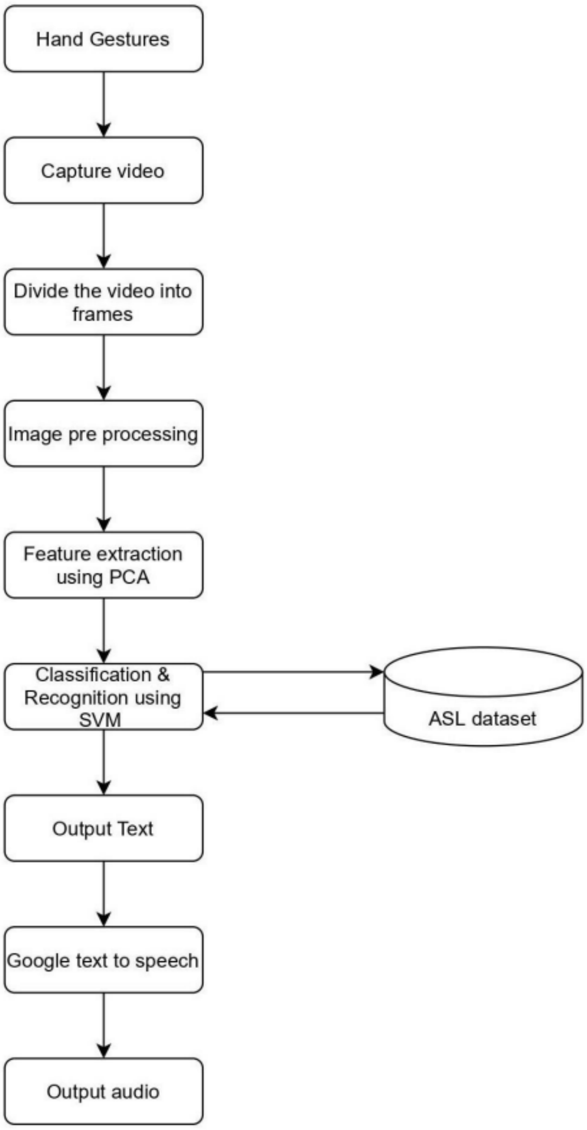

Figure 4.5: SVM

4.5 Real Time Prediction

The model takes a live sequence of video frames from the webcam, it fixes the background of our roi, extracts the hand region, and finally outputs a threshold image. This threshold image is given to our trained model as an input. The image is processed and our model gives two outputs - predictedClass , confidence.We take the maximum confidence class as the predictedClass. As we have created 32 distinct gestures so the model contains 32 predictedClasses consisting of alphabets, special characters and custom gestures.For improved communication, we have added the text-mode which makes it simple to convey messages between people using signs. We have also added the text to sound feature to make conversations easier.

4.6 Voice Conversion

After predicting the hand gestures the text will be converted to speech using Google Text-to-Speech. gTTS (Google Text-to-Speech)is a Python library and CLI tool to interface with Google Translate text-to-speech API. We will import the gTTS library from the gtts module which can be used for speech translation. The text variable is a string used to store the user's input.

Chapter 5

Results and Discussion

5.1 Dataset Description

We have created our own Dataset by following the American Sign Language for alphabet, special characters and also added a few custom gestures like YOLO , All the best, etc.. which can be further extended to add Indian Sign Language Vocabulary. We first segment the Hand region by using Background Subtraction Technique which uses the concept of running averages from a sequence of video frames.Then we do thresholding for this image. By this we can get the contour of the hand region.For each gesture we get a threshold image and all these images are pre-processed using Keras preprocessing model which does the necessary resizing, shifts, flips, orientations to a certain degree, etc.. which concludes our Final Dataset.

Figure 5.1: Dataset images

5.2 Parameter Setup

Training parameter	CNN
Learning Rate	0.001
Batch Size	64
Optimizer	Adam
Loss Function	Categorical cross entropy
Epochs	50

Figure 5.2: CNN Parameter Setup

5.3 Results

Figure 5.3: Web page

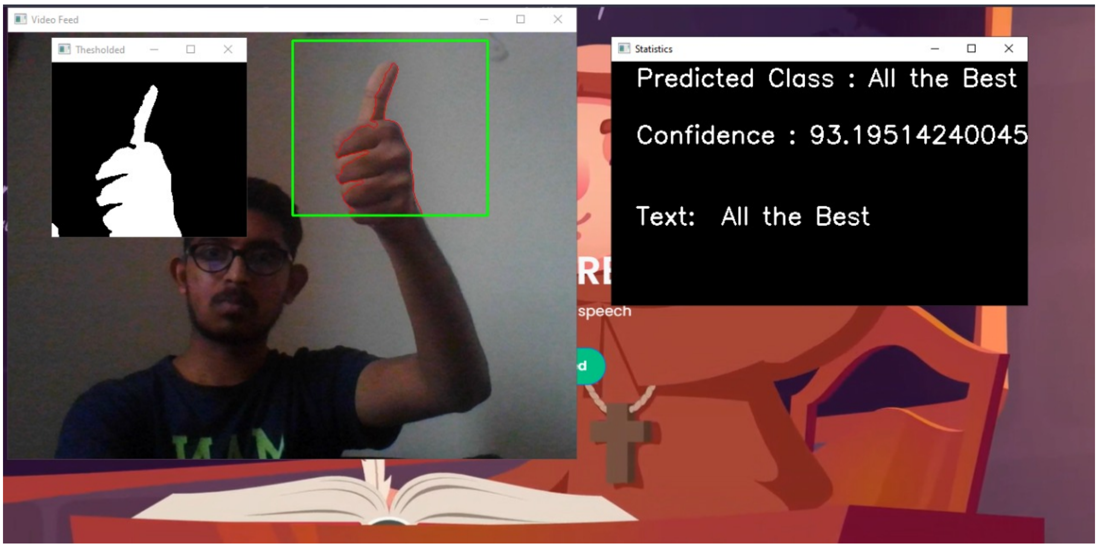

Figure 5.4: Sign language recognition

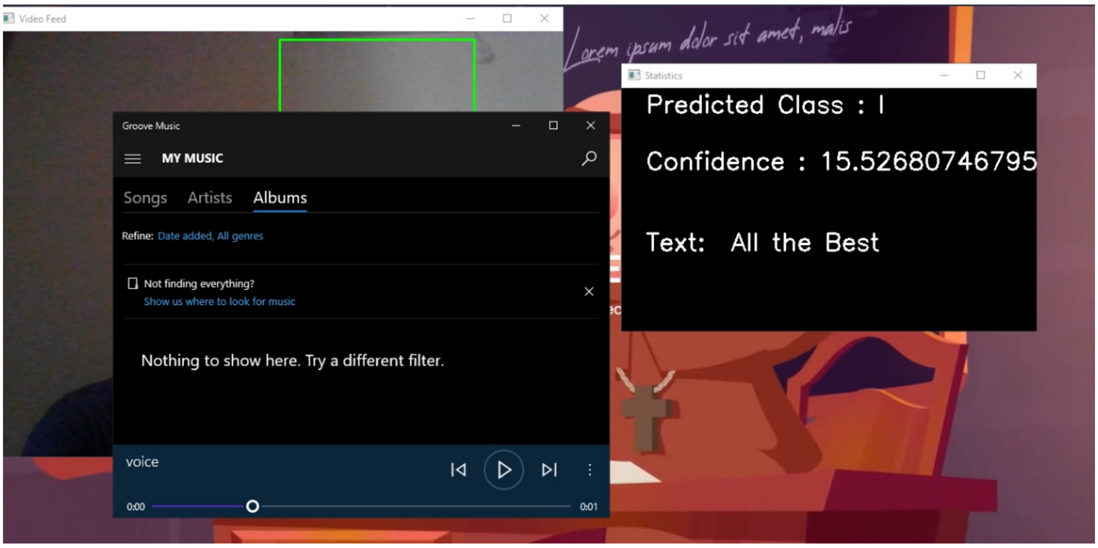

Figure 5.5: Speech prediction

Models	Training Accuracy	Testing Accuracy
CNN	99.93	99.64
SVM	91.57	84.18

Figure 5.6: Accuracy

```
# Train model
model.fit(loadedImages, outputVectors, n_epoch=50,
          validation_set = (testImages, testLabels),
          snapshot_step=100, show_metric=True, run_id='convnet_coursera')

model.save("TrainedModel/GestureRecogModel.tfl")
```

```
Training Step: 4299  | total loss: 0.00507 | time: 52.982s
| Adam | epoch: 050 | loss: 0.00507 - acc: 0.9992 -- iter: 5440/5504
Training Step: 4300  | total loss: 0.00460 | time: 61.361s
| Adam | epoch: 050 | loss: 0.00460 - acc: 0.9993 | val_loss: 0.01416 - val_acc: 0.9964 -- iter: 5504/5504
--
INFO:tensorflow:C:\Users\Rohit\Desktop\Hand Gesture Recognition f\Hand Gesture Recognition\Mydataset\TrainedMode
ogModel.tfl is not in all_model_checkpoint_paths. Manually adding it.
```

Figure 5.7: Result

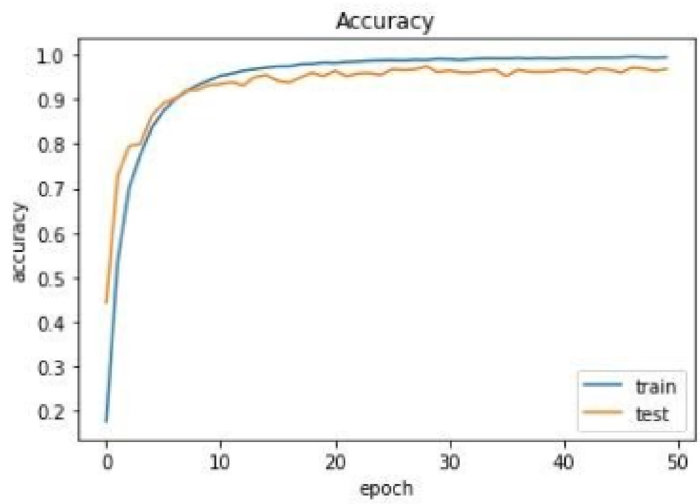

Figure 5.8: CNN Accuracy

Chapter 6

Conclusion and Future Work

Some of the limitations of our model are the model requires good lighting conditions and as background subtraction technique was used so we need to make sure that the background is stable otherwise the prediction may go wrong.As we are limited to make only a few set of signs because of using hands so maybe we cannot cover all ranges of signs. As we use text mode the size of the text is limited so we cannot form large messages at once. So we can improve the model in future to extend its effectiveness further and eliminate the limitations.

Chapter 7

References

1. Sunitha K. A, Anitha Saraswathi.P, Aarthi.M, Jayapriya. K, Lingam Sunny, "Deaf Mute Communication Interpreter- A Review", International Journal of Applied Engineering Research ,Volume 11, pp 290-296 ,2016.

2. Madhavan Suresh Anand, Nagarajan Mohan Kumar, Angappan Kumaresan, " An Efficient Framework for Indian SignLanguage Recognition Using Wavelet Transform" Circuits and Systems, Volume 7, pp 1874-1883, 2016.

3. Mandeep Kaur Ahuja, Amardeep Singh, "Hand Gesture Recognition Using PCA", International Journal of Computer Science Engineering and Technology (IJCSET), Volume 5, Issue 7, pp. 267-27, July 2015.

4. Sagar P.More, Prof. Abdul Sattar, "Hand gesture recognition system for dumb people"

5. International Journal of Science and Research (IJSR)

6. Chandandeep Kaur, Nivit Gill, "An Automated System for Indian Sign Language Recognition",International Journal of Advanced Research in Computer Science and Software Engineering.

7. Pratibha Pandey, Vinay Jain, "Hand Gesture Recognition for Sign Language Recognition: A Review",International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 3,March 2015 .

8. Nakul Nagpal,Dr. Arun Mitra.,Dr. Pankaj Agrawal, "Design Issue and Proposed Implementation of Communication Aid for Deaf Dumb People", International Jour-

nal on Recent and Innovation Trends in Computing and Communication ,Volume: 3 Issue: 5,pp- 147 – 149.

9. Neelam K. Gilorkar, Manisha M. Ingle, "Real Time Detection And Recognition Of Indian And American Sign Language Using Sift", International Journal of Electronics and Communication Engineering Technology (IJECET), Volume 5, Issue 5, pp. 11-18 , May 2014

10. Neelam K. Gilorkar, Manisha M. Ingle, "A Review on Feature Extraction for Indian and American Sign Language", International Journal of Computer Science and Information Technologies, Volume 5 (1) , pp- 314-318, 2014.

11. Ashish Sethi, Hemanth ,Kuldeep Kumar,Bhaskara Rao ,Krishnan R, "Sign Pro-An Application Suite for Deaf and Dumb", IJCSET , Volume 2, Issue 5, pp-1203-1206, May 2012.

12. Priyanka Sharma,"Offline Signature Verification Using Surf Feature Extraction and Neural Networks Approach", International Journal of Computer Science and Information Technologies, Volume 5 (3) , pp 3539-3541, 2014.

13. F. K. H.Quek, "Toward a Vision-Based Hand Gesture Interface," in Virtual Reality Software and Technology Conference, (1994) 17-31.

14. F. K. H. Quek, "Eyes in the Interface," Image and Vision Computing, 13 (1995).

15. R. A. Bolt, "Put-that-there: Voice and Gesture at the Graphics Interface," Proc. SIGGRAPH80, (1980).

16. T. Starner and A. Pentland, "Real-Time American Sign Language Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence 20, (1998) 1371–1375.

17. C. W. Ng and S. Ranganath, "Real-Time Gesture Recognition System and Application," Image Vis. Comput.,20(13–14) (2002) 993–1007.

18. K. Abe, H. Saito, and S. Ozawa, "Virtual 3-D Interface System via Hand Motion Recognition from Two Cameras,"IEEE Trans. Syst., Man, Cybern. A, 32(4) (2002) 536–540.

19. Md. Hasanuzzaman, V. Ampornaramveth, Tao Zhang, M.A. Bhuiyan , Y. Shirai and H. Ueno, "Real-Time Vision-Based Gesture Recognition for Human Robot Interaction", in the Proc. of IEEE Int. Conf. on Robotics and Biomimetics, Shenyang China, (2004).

20. Vuthichai Ampornaramveth and Haruki Ueno, "Software Platform for Symbiotic Operations of Human and Networked Robots", NII Journal, 3 (2001) 73-81.

21. J. Segen, "Controlling Computers with Gloveless Gestures," in Proc. Virtual Reality Systems Conf., New York, (1993)66–78.

22. D. Kortenkamp, E. Huber, and R. P. Bonasso, "Recognizing and Interpreting Gestures on a Mobile Robot," in Proc. Artificial Intelligence (AAAI) Workshop, Portland, (1996) 915–921.

23. Elena Sanchez-Nielsen, Luis Anton-Canalís and, Mario Hernández-Tejera , "Hand Gesture Recognition for Human-Machine Interaction", Journal of WSCG, 12(1-3) (2003).