Hand Gesture Recognition To Speech Conversion

Dhruv Vaghela[1] , Ankit Vaity[2] , Rajvi Makwana[3] , Afreen Shaikh[4], Chanda Chouhan[5]

1(Department of Information technology, Mumbai University, India)
2(Department of Information technology, Mumbai University, India)
3(Department of Information technology, Mumbai University, India)
4(Department of Information technology, Mumbai University, India)

Abstract: *Inability to speak is considered to be true disability. In our system we intend to overcome this disability by capturing hand gestures performed by the disabled and giving text and speech as output. This will not only help the mute people but also deaf and blind ones to communicate with anyone around. Our system will contain a manual of our own gestures which will be categorized depending on the situation they are most likely to be used in hence making it easier for the user to convey his message.*

I. Introduction

The motive of the paper is to provide medium for the disabled and help them communicate. Since gestures play a very important role thus the usage of gesture recognition can very be applied in various sections of the society, as its application ranges from the scientific usage, Human Computer Interactions (HCI) and even for the revivification of society. Sign Language is an abstract entity directed at natural language whose origin depends on the "Sign" or the "Gestures" and is the natural way for communication between the taciturn and vocally-debilitated people[1]. Our system will recognize the hand gestures present in the manual and give text as well as speech as output which will help the blind, deaf and dumb to communicate with each other without the help of human mediator. The main paradigm that we focus on is to endeavor the linking between the Sign Language medium with the Standard English Language and thus providing the communication between the two communities in a seamless experience. The main concept simply focuses on the gesture capturing using mobile computing device. Initially the gesture which is to be captured is added in the Training Database and the respective word is being stored. When the gesture is being displayed in front of the inbuilt camera of the mobile device, it captures and attempts to map with the gesture that is stored in the Training Database, the mapping procedure of the gesture is done using CNN and various Open CV algorithms. After the particular image has been mapped, by matching the contours of the stored image in the database, the respective word is being triggered and the finally the speech to voice conversion is performed.[1]

II. Requirement Analysis

Tensorflow.js: TensorFlow.js is an open source WebGL-accelerated JavaScript library for machine intelligence. It brings highly performant machine learning building blocks to your fingertips, allowing you to train neural networks in a browser or run pre-trained models in inference mode.

Keras: Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear and actionable feedback upon user error. This makes Keras easy to learn and easy to use. As a Keras user, you are more productive, allowing you to try more ideas than your competition, faster -- which in turn helps you win machine learning competitions.

Python3: Python language is one of the most flexible languages and can be used for various purposes. Python has gained huge popularity base of this. Python does contain special libraries for machine learning namely scipy and numpy which great for linear algebra and getting to know kernel methods of machine learning. The language is great to use when working with machine learning algorithms and has easy syntax relatively. For beginners, this is the best language to use and to start with.

Heroku: Heroku is a cloud platform as a service. That means you do not have to worry about infrastructure; you just focus on your application. Heroku provides very well written tutorial which allows you to start in minutes. Also they provide first 750 computation hours free of charge which means you can have one processes (aka

Dyno) at no cost. Also performance is very good e.g. simple web application written in flask/django can handle around 60 - 70 requests per second.

Webcam: A webcam is a video camera that feeds or streams its image in real time to or through a computer to a computer network. When "captured" by the computer, the video stream may be saved, viewed or sent on to other networks travelling through systems such as the internet, and e-mailed as an attachment. When sent to a remote location, the video stream may be saved, viewed or on sent there. Unlike an IP camera (which connects using Ethernet or Wi-Fi), a webcam is generally connected by a USB cable, or similar cable, or built into computer hardware, such as laptops.

III. Figures And Tables

Figure no 1: System Block Diagram

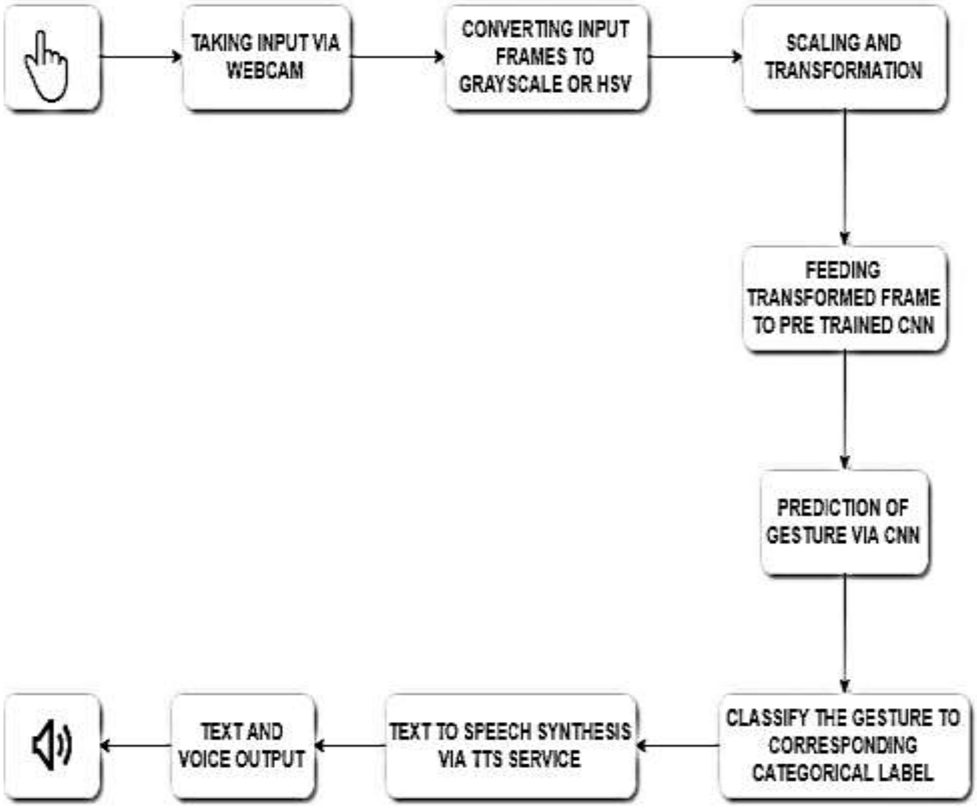

TAKING INPUT VIA WEBCAM
CONVERTING INPUT FRAMES TO GRAYSCALE OR HSV
SCALING AND TRANSFORMATION
FEEDING TRANSFORMED FRAME TO PRE TRAINED CNN
PREDICTION OF GESTURE VIA CNN
CLASSIFY THE GESTURE TO CORRESPONDING CATEGORICAL LABEL
TEXT TO SPEECH SYNTHESIS VIA TTS SERVICE
TEXT AND VOICE OUTPUT

Figure no 2: Sequence diagram

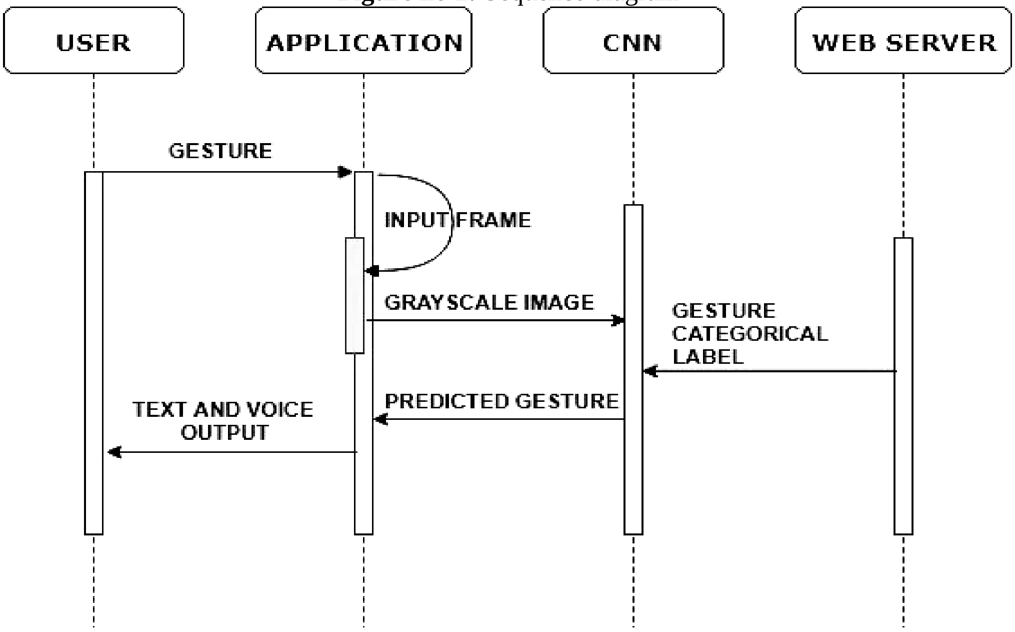
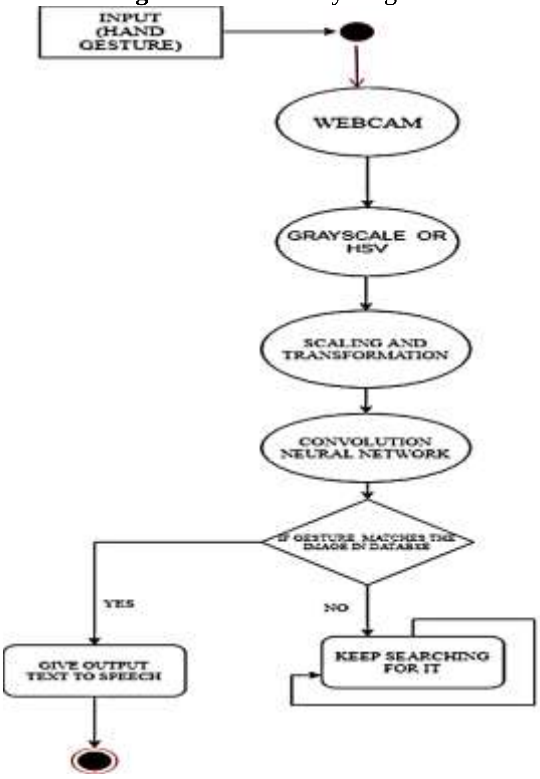


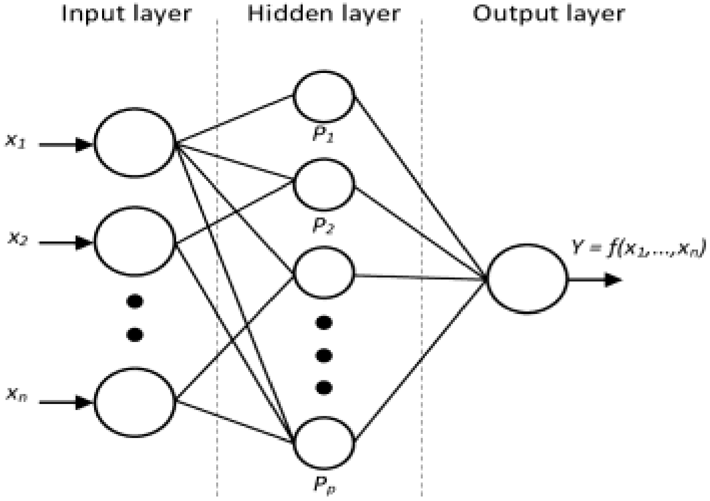
Figure no 3: Activity diagram


INPUT (HAND GESTURE)
WEBCAM
GRAYSCALE OR HSV
SCALING AND TRANSFORMATION
CONVOLUTION NEURAL NETWORK
IF GESTURE MATCHES THE IMAGE IN DATABASE
YES
NO
GIVE OUTPUT TEXT TO SPEECH
KEEP SEARCHING FOR IT

IV. Proposed System

Using hand wearable devices is not as efficient as using a web app to recognize gestures as they increase the hardware cost. The basic idea of our web app is to use inbuilt web cam to capture gestures. Initially it captures images in RGB format and then the image is scaled and transformed accordingly. The system works on convolutional neural network (CNN). The first layers that receive an input signal are called convolution filters. Convolution is a process where the network tries to label the input signal by referring to what it has learned in the past. If the input signal looks like previous "Hello" gesture images it has seen before, the "Hello" reference signal will be mixed into, or convolved with, the input signal. The resulting output signal is then passed on to the next layer. The output signal strength is not dependent on where the features are located, but simply whether the features are present. Inputs from the convolution layer can be "smoothened" to reduce the sensitivity of the filters to noise and variations. This smoothing process is called subsampling, and can be achieved by taking averages or taking the maximum over a sample of the signal. Examples of subsampling methods (for image signals) include reducing the size of the image, or reducing the color contrast across red, green, blue (RGB) channels. The next step is Activation. The activation layer controls how the signal flows from one layer to the next, emulating how neurons are fired in our brain. Output signals which are strongly associated with past references would activate more neurons, enabling signals to be propagated more efficiently for identification. CNN is compatible with a wide variety of complex activation functions to model signal propagation, the most common function being the Rectified Linear Unit (ReLU), which is favored for its faster training speed. The last layers in the network are fully connected, meaning that neurons of preceding layers are connected to every neuron in subsequent layers. This mimics high level reasoning where all possible pathways from the input to output are considered. When training the neural network, there is additional layer called the loss layer. This layer provides feedback to the neural network on whether it identified inputs correctly, and if not, how far off its guesses were. This helps to guide the neural network to reinforce the right concepts as it trains. This is always the last layer during training.

Figure no 3: Mathematical Model



The image that is captured through the webcam is initially passed and processed through the Original Image filter, due to which it processed into required format and according to the threshold values the image is being detected and the respective word and the speech is displayed and spoken.

V. Future Scope

The advancement of the system initially started from hand-gloves and sensors. The drawback of this approach is that we need to rely on hardware devices which are complex and not easily affordable. The technology introduced by us is free from such hardware devices as it needs only an inbuilt system camera. The only drawback the system faces is the amount of time it takes to train the model. The future scope of this system can be a system which is used to detect objects in front of blind people with the help of voice synthesizer. It can inform them about the huddles in their path while walking on the streets alone.

VI. Conclusion

In this paper, we presented the concept of gesture-to-speech conversion concept, due to which the communication between the vocally impaired people of the society and the common people will be carried out without any obstruction. The explanation of the design and implementation is presented along with the prototype which captures the gesture of the ARDA sign language. As compared to the other system this concept not only focuses on the gesture to word display but also on the speech synthesis.

Acknowledgment

The success and outcome of this project required a lot of guidance and assistance for many people and we are extremely privileged to have got this all along the completion of our project. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

We owe our deep gratitude to our project guides Prof. Chanda Chauhan, who took keen interest in our project and guided in all along, till the completion of our project work by providing all the necessary information for developing a good system.

We are thankful and fortunate to get constant encouragement, support and guidance from all the teaching staff of Bachelor of Engineering in Information Technology of Atharva College of Engineering, which helped us a lot in successfully completing our project work.

Lastly, I would like to express my appreciation towards my fellow classmates and friends for providing us the moral support and encouragement

References

[1]. Hand-Gesture Recognition for Automated Speech Generation. IJSRD|Vol. 4,Issue 02, 2016.

[2]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "ImageNet: A large-scale hierarchical image database", CVPR, 2009.

[3]. A. Howard, "Some improvements on deep convolutional neural network based image classification", ICLR, 2014.

[4]. J. Redmon, A. Angelova, "Real-time grasp detection using convolutional neural networks", IEEE International Conference on Robotics and Automation, pp. 1316-1322, 2015.