IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731

Artificial-Intelligence Virtual Assistant for People With Hearing Loss and Speech Impairment

Anup Maurya[1], Pratik Pawade[2], Shashank Shastri[3], Bhoomik Kini4, Namrata Jaiswal[5]
Project Guide, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
Vishwatmak Om Gurudev College of Engineering, Aghai, Maharashtra, India

Abstract: *When you see a gesture, your brain can easily tell what the image is about, but can a computer tell what the gesture is representing? Computer vision researchers worked on this a lot and they considered it impossible until now! With the advancement in Deep learning techniques and availability of huge datasets and computer power, we can build models that can recognize what a gesture means. Voice assistants are intelligent personal assistants that can help someone with basic tasks. They often understand natural language and can help with things like creating meeting requests, reporting a sports score, and sharing the weather forecast. Intelligent personal assistants have access to a large amount of information on a device or online, which enables them to perform simple tasks. These virtual assistants though are limited to people who can speak. People with hearing loss and speech impairment have no access to such devices and thus are not availing these benefits. Now combining virtual assistant with gesture recognition we can create such a virtual assistant that will not use voice as a medium but sign language. This will remove the existing barrier for people with hearing loss and speech impairment and will give them access to vast amounts of information.*

Keywords: Indian Sign Language; DWT; Computer Vision; Convolutional Neural Network (CNN); Recurrent Neural Network (RNN); natural language processing; avatar; sign movement; Gesture recognition; Sign language recognition.

I. Introduction

Voice assistants are intelligent personal assistants that can help someone with basic tasks. They often understand natural language and can help with things like creating meeting requests, reporting a sports score, and sharing the weather forecast. Intelligent personal assistants have access to a large amount of information on a device or online, which enables them to perform simple tasks. As the name implies, it relies on voice as the input, which therefore is out of reach for people with hearing loss and speech impairment.

Any non-verbal communication like the motion of hands, face, and other body parts is a form of gesture. Gesture recognition enables computers to understand human actions. Gesture recognition algorithm acts as an interpreter between computer and human. This could provide the potential for humans to interact naturally with computers without any physical contact with mechanical devices. Gestures are performed by people with hearing loss and speech impairment to communicate via sign language.

Combining these 2 concepts, we can create a virtual assistant app capable of carrying out commands based on signs instead of voice. This removes the barrier and allows people with hearing loss and speech impairment to experience virtual assistants. The response from the virtual assistant is converted from text to signing via animation.

We chose to make an android app as an android device that would have a front camera for the gesture input, sufficient processing power on the fly, and is the widely used OS.

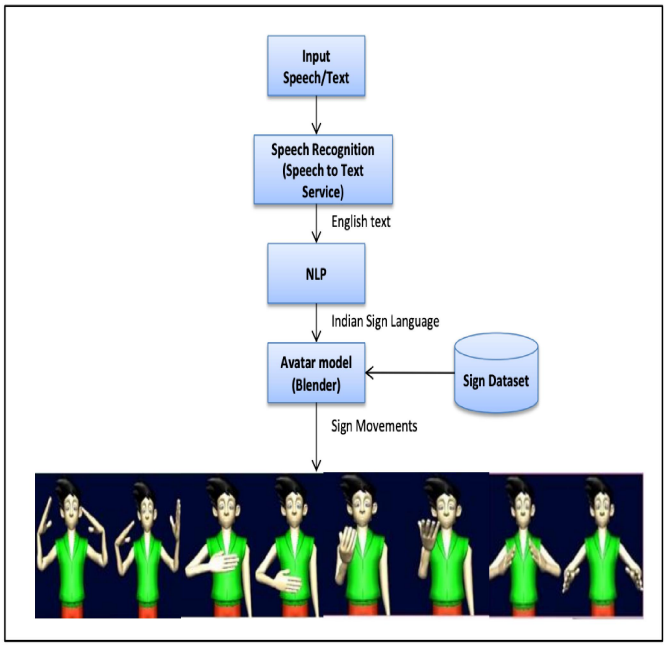
II. Literature Survey

Our research of existing systems shows the only project close to what we propose is a gesture recognition system which needs initial training before use and it uses a text-to-speech system to speak the interpreted sign to Alexa and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731

then a speech to text system to transcribe the response from Alexa for the user. However, this requires an Alexa device or a google assistant device, a laptop/desktop, and the initial training.

In [4] Abhishek uses TensorFlow.js which allows quick prototyping and to run models directly in the browser. This is huge, from a standpoint of portability, speed of development, and ability to easily interact with web interfaces. Also, the models run entirely in the browser without the need to send data to a server. He uses a kNN model which performs faster/better than CNN's on smaller datasets. They become memory intensive and performance drops when training with many examples but since he knew his dataset would be small this was a non-issue. Since kNNs aren't learning from examples, they are poor at generalization. Thus the prediction of a model trained on examples made up entirely of one person will not transfer well to another person. This again was a non-issue for him since he would be both training and testing the model by repeatedly performing the signs himself. Though this does not serve well for our purpose as it has many disadvantages for us, one being initial training needs to be done by users, poor at generalization, as we'll be dealing with a large dataset, it would become memory intensive and performance drops. In [5] Debashis et al. describe how a speech/text input can be translated to Indian Sign Language via a 3D avatar.



III. Proposed System

This project focuses on the proposed continuous ISL gesture recognition system. Dataset consists of a collection of signs where the single hand or both the hands have been used for performing continuous ISL gestures. Every sentence is a combination of static and dynamic gestures. Extracting the start frame and end frame of each gesture is the main problem in a continuous sign language gesture recognition system because it consists of a collection of meaningful gestures and also vague gestures having no meaning.

We deal with this problem using a gradient-based key frame extraction method. Here a major change in the gradient shows the end of one gesture and the start of another gesture. The keyframe helps to break each sentence into a sequence of words (isolated gestures) and is also obligatory for extracting frames of meaningful gestures. Orientation histogram, DWT, and PCAare used for extracting features of those frames which comprises meaningful gestures. The general diagram of gesture recognition is shown in the framework analysis.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731

IV. FRAMEWORK ANALYSIS

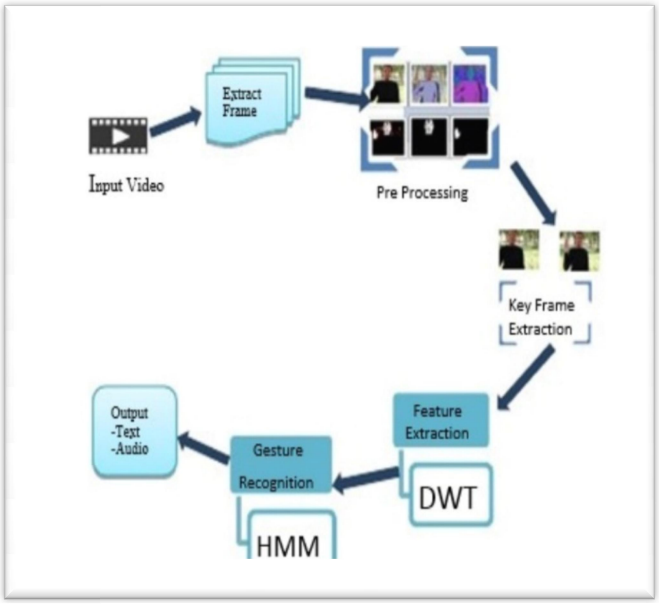


Figure: General Diagram of Proposed Framework

This diagram is the general purpose of framework analysis. First, take the input using image processing algorithms and the detection algorithm captures images and videos. Then extract these frames and use pre-processing. Then move the next step keyframe extraction this step is used in the removal of the double capture images. Feature extraction involves simplifying the number of resources required to describe a large set of data accurately. DWT can be used for high dimensionality data analyses, such as image processing and image data analysis.

A method based on Hidden Markov Models (HMMs) is presented for dynamic gesture trajectory modeling and recognition. Adaboost algorithm is used to detect the user's hand and a contour-based hand tracker is formed combining condensation and partitioned sampling. The response from virtual assistant API is converted from text into an animation of Indian Sign Language.

V. ALGORITHMS

5.1 Image Acquisition

Image acquisition is the first step in any vision system, only after this process, you can go forward with the image processing. This application is done by using the IPWebCam android application. The application uses the camera present in the phone for continuous image capturing and a simultaneous display on the screen. The image captured by the application is streamed over its Wi-Fi connection (or WLAN without internet as used here) for remote viewing. The program accesses the image by logging into the device's IP, which is then shown in the GUI.

5.2 Image Pre-Processing: Edge Detection

In this program, the edge detection technique used is the Sobel edge detector. The image captured is then passed through a Sobel filter.

5.3 Thinning

Thinning is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion or opening. It can be used for several applications but is particularly useful for skeletonization. In this mode, it is commonly used to tidy up the output of edge detectors by reducing all lines to single-pixel thickness. Thinning is normally only applied to binary images, and produces another binary image as output. After the edge detection, thinning has to be performed. Thinning is applied to reduce the width of an edge to a single line.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731

5.4 Hand Token

The idea here is to make the image into a neuronal network usable form so that the cosine and sine angles of the shape represent the criteria of a recognition pattern. Each square represents a point on the shape of the hand image from which a line to the next square is drawn.

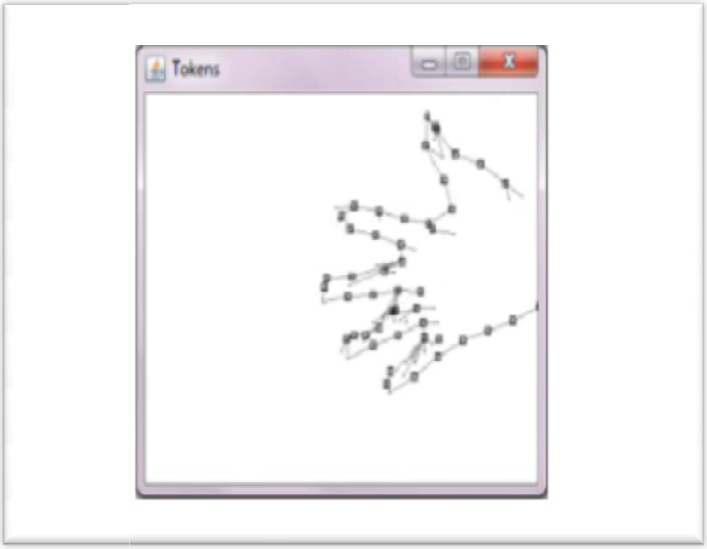


Figure: Generated token of the original image

On zooming a part of figure 5 it shows a right-angled triangle between the 2 consecutive squares. This and the summary of all triangles of a hand image are the representation of the tokens of a hand from which we can start the neuronal network calculations.

The right-angled triangle in figure 5 represents a token of a single hand image. The angles A and B are the two necessary parts that will be fit into the neuronal network layers. With the two angles, we can exactly represent the direction of the hypotenuse from point P1 to P2 which represents the direction of a hand image.

VI. Requirement Analysis

6.1 Software Requirement

- Android Studio.
- Alexa Voice Service SDK.
- Tensor flow.
- Blender.

6.2 Hardware Requirement

- Computer/Laptop With Dedicated Graphics Card.
- Android Smartphone

VII. Design Analysis

This figure show workflow and the design of our project application. In the Below steps.

- The user shows a hand sign which is captured by a camera on the mobile phone.
- The hand sign is recognized and converted to text.
- The text is sent to a virtual assistant API.
- The response from API is converted to sign language.
- Sign language is shown as an animation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731

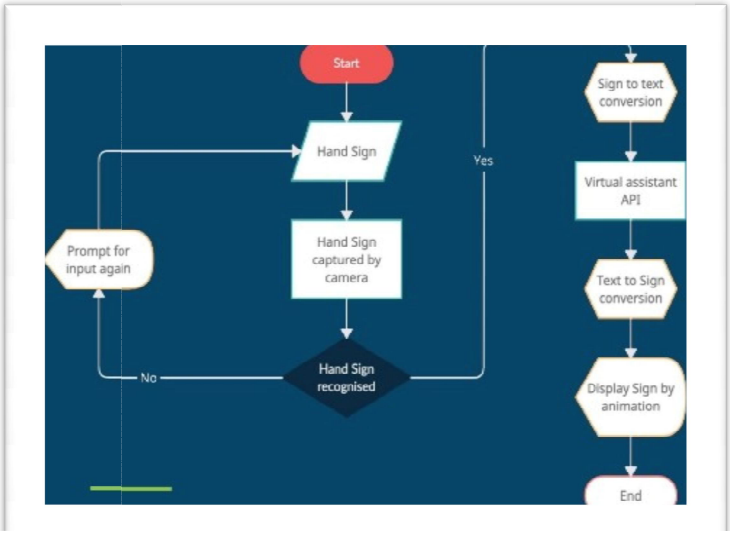
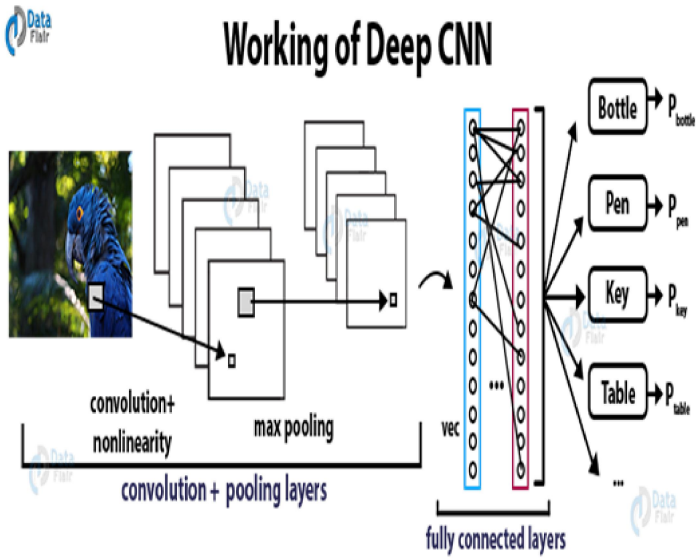


Figure: Flow Chart of Design Our Project

VIII. PROJECT ARCHITECTURE

7.1 CNN (Convolutional Neural Networks)

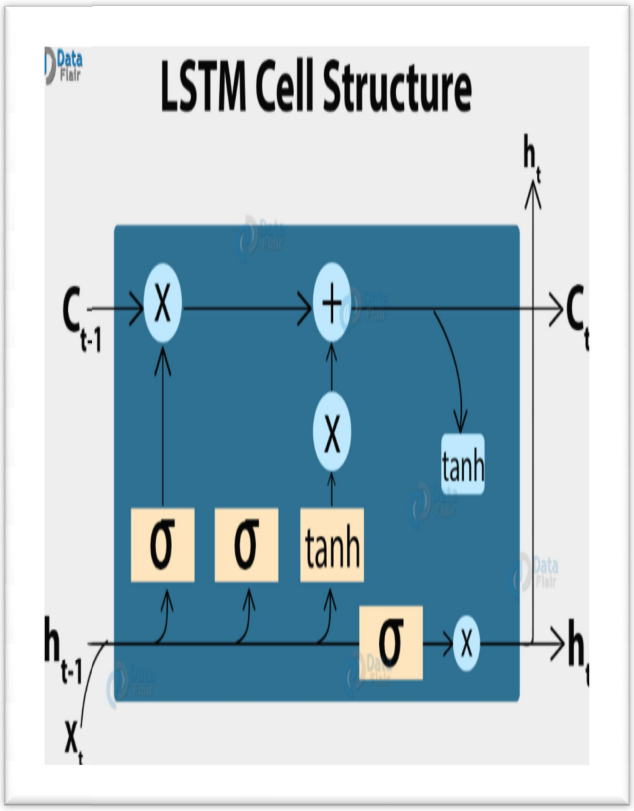


Convolutional Neural networks are specialized deep neural networks that can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images. CNN is used for image classifications and identifying if an image is a bird, a plane, or Superman, etc. It scans images from left to right and top to bottom to pull out important features from the image and combines the features to classify images. It can handle the images that have been translated, rotated, scaled, and changes in perspective.

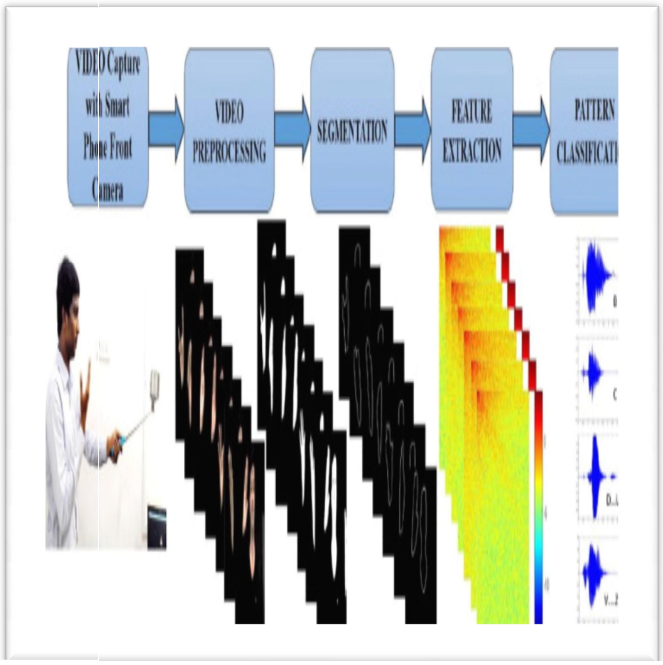
7.2 LSTM (Long Short Term Memory)

LSTM stands for Long short-term memory, they are a type of RNN (recurrent neural network) that is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short-term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731



7.3 Sign Language Recognition Model



So, to make the sign language recognition model, we have merged these architectures. It is also called a CNN-RNN model. CNN is used for extracting features from the image. We will use the pre-trained model. LSTM will use the information from CNN to help generate a proper sentence from the words recognized.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 12, Issue 1, December 2021
Impact Factor: 5.731

IX. METHODOLOGY

9.1 Data Collection

The first step is to collect and or create a dataset of Indian Sign language, to create a prediction model from the dataset in Python, and to make it in the format of TensorFlow lite so that it can be easily used in the android app.

9.2 Model Creation

Creating a prediction model based on the dataset by splitting data into a training dataset and testing dataset.

9.3 Creating an Android App

The android app will use the mobile phone's camera to record the signing and use the pre-trained model to predict what the sign means.
Using the continuous ISL gesture recognition system we recognize the hand sign shown via an android camera.
Using LSTM we infer the correct sentence from the gestures.
Using the Alexa Voice Service SDK for android we process the command.
Using the Natural Language Processing (NLP) technique the text response is converted to the 3D avatar.

X. CONCLUSION

The Proposed framework for continuous ISL gesture gives satisfactory performance including features obtained using discrete wavelet transform where both the hands have been used for performing any gesture. If the gesture recognition system cannot give results in a real-world environment, then it is useless. Here our system performs correctly in any type of background environment. This makes my system more flexible. In a continuous ISL recognition system, keyframe extraction is the foremost step. It helps us to extract isolated gestures from continuous gestures. It also shows how many frames an isolated gesture will have. It decreases the training and testing time. After hand segmentation, we applied DWT for extracting features of hands for the training dataset as well as for testing. Here classification accuracy is measured with the maximum number of matched frames. Experimental results show that the proposed method gives satisfactory results with the hidden Markov model. Results are also tested using the normal webcam and get appropriate results. This work has been enhanced by creating a dataset with different backgrounds and different illumination conditions. Here we applied more appropriate features which incorporate the shape of the hand in the time of acting gestures, speed of performing each gesture, etc. There are various other classifiers like a k-Nearest neighbor, Support vector machine (SVM) that have been applied for classification.

REFERENCES

[1]. Yi Li, Weidong Chen, Yang Zheng, "Dynamic hand gesture recognition using hidden Markov models", IEEE Computer Science & Education (ICCSE), 7th International Conference, pp. 360 - 365, 14-17 July 2012.
[2]. Deng J.W., Tsui H.T., "An HMM-based approach for gesture segmentation and recognition", IEEE Pattern Recognition, 2000. Proceedings. 15th International Conference, Vol. 3, pp. 679 - 682, 2000.
[3]. Starner, T., Pentland, A., "Real-time American Sign Language recognition from video using hidden Markov models", Computer Vision, 1995. Proceedings, International Symposium on, pp. 265-270, 21-23 Nov 1995, DOI. 10.1109/ISCV.1995.477012.
[4]. Getting Alexa to Respond to Sign Language Using Your Webcam and TensorFlow.js — The TensorFlow Blog.
[5]. Debashis Das Chakladar, Pradeep Kumar, Shubham Mandal, ParthaPratim Roy, Masakazu Iwamura, Byung-Gyu Kim, "3D Avatar Approach for Continuous Sign Movement Using Speech/Text", Appl. Sci. 2021, 11, 3439. https://doi.org/10.3390/app11083439