Realtime Hand Gesture Recognition using LSTM model and Conversion into Speech

Sakshi Mankar[1], Kanishka Mohapatra[2], Ashwin Avate[3], Mansi Talavadekar[4], Prof. Surendra Sutar[5]
[1,2,3,4]*Student, MCT Rajiv Gandhi Institute of Technology, Mumbai*
[5]*Faculty, MCT Rajiv Gandhi Institute of Technology, Mumbai*

***Abstract*— People require communication to communicate with each other. "Specially abled people", those who have speech or hearing disorder, "Mute" and "Deaf" people respectively, are always dependent on some sort of visual communication. People without visual and hearing disabilities sometimes face difficulties and cannot communicate with specially abled people due to lack of sign language education. Sign language is well received among them and they use it to express themselves. To achieve two-way communication between specially abled people and the general public there is a need to build a system that can interpret the gestures into text and speech. A vision-based technology of hand gesture recognition is an important part of human-computer interaction. Technology like gesture recognition can help us build a framework that can interpret sign language/gesture into text and speech. Gestures by hand which can represent a notion using unique shapes and finger position have a scope for human machine interaction. The major steps involved in designing the system are: tracking, segmentation, gesture acquisition, feature extraction, gesture recognition and conversion into speech.**

***Index Terms*— Gestures, LSTM neural network, ReLU activation function, sign-language, TensorFlow.**

I. INTRODUCTION

A crucial application of gesture recognition is sign language detection. Current technologies for gesture recognition can be divided into two types: sensor-based and vision-based. In sensor-based methods, data glove or motion sensors are incorporated from which the data of gestures can be extracted. Even minute details of the gesture can be captured by the data capturing glove which ultimately enhances the performance of the system. However, this method requires wearing a data capturing hand glove with embedded sensors that makes it a bulky device to carry. This method affects the signer's usual signing ability and it also reduces user amenity. Vision based methods include image processing. This approach provides a comfortable experience to the user. The image is captured with the help of cameras. No extra devices are needed in the vision-based approach. This method deals with the attributes of the image such as color and texture that are obligatory for integrating the gesture. Although the vision-based approach is straightforward, it has many challenges such as the complexity and convolution of the background, variations in illumination and tracking other postures along with the hand object, etc. arise. Sign language provides a way for speech impaired and hearing-impaired people to communicate with other people. Instead of a voice, sign language uses gestures to communicate. Sign language is a standardized way of communication in which every word and alphabet is assigned to a distinct gesture. It would be a win-win situation for both specially abled people and the general public if such a system is developed where sign language could be converted into text/speech. Technology is advancing day after day but no significant improvements are undertaken for the betterment of specially-abled communities. About nine million people in the world are deaf and mute. Communication between specially abled people and general people have always been a challenging task but sign language helps them to communicate with other people. But not everyone understands sign language and here is where our system will come into the picture.

II. LITERATURE SURVEY

[1] They propose a fast-multi-scale feature detection and description method for hand gesture recognition.

Where they firstly approximate complex Gaussian derivatives with simple integral images in feature detection. Then multiscale geometric descriptors at feature points are obtained to represent hand gestures. Finally, the gesture is recognized with its geometric configuration. Specific gesture is required to trigger the hand detection followed by tracking; the hand is further segmented using motion and color cues. The scale-space feature detection is integrated into gesture recognition.

[2] In this paper they have mentioned two algorithms on which the TTS system works which is character to voice and word pronunciation. Further they worked on the problems in the TTS system mentioned as text processing, text-to-phonetic conversion and pronunciation. As they have seen, that delay in sound waves causes the speech to look very unnatural. To overcome this problem, they mentioned a method which recognizes the delay and automatically removes it. And the best method would be 'integration of synthesized speech'.

[3] In their work, they build on the results of an existing project. The project proposed a CNN model that recognized a set of 50 different signs in the Flemish Sign Language with an error of 2.5%, using Microsoft Kinect. Unfortunately, this work was limited in the sense that it considers only a single person in a fixed environment. Lionel Pigou and teammates used the data set from the ChaLearn Looking at People 2014 consisting of 20 different Italian gestures, performed on 27 users with variations in surroundings, clothing, lighting and gesture movements. They used a max-pooling method in their architecture consisting of two CNNs, one for extracting hand features and other for upper body features. The model learning rate is initialized at 0.003 with a 5% decreasing each epoch. The weights of the CNNs are randomly initialized with a normal distribution with $\mu = 0$ and $\sigma = 0.04$, and $\sigma = 0.02$ for the weights of the ANN. The biases of CNN and ANN here are initialized at 0.2 and 0.1 respectively. The accuracy on the testing set is 95.68% and observed a 4.13% false positive rate, caused by the noise movements.

[4] This paper focuses on designing a vision-based hand gesture recognition system with a high correct detection rate along with a high-performance, which can work in a real time Human Computer Interaction system without having any limitations (gloves, uniform background, etc.) on the user environment. Input to the system is a pre-recorded video sequence which later it detects the skin color by using an adaptive algorithm. For the current user skin color has to be fixed based on the lighting and camera parameters and condition. Once it has been fixed, the hand is localized with a histogram clustering method. Then a machine learning algorithm is used to detect the hand gestures in consecutive frames to distinguish the current gesture. This paper describes how the execution of the system is done based upon the images captured. Hand detection is done using OpenCV and TensorFlow object detectors. And further it is enhanced for analysis of gestures by the computer to perform actions like switching the pages, scrolling up or down the page.

[5] This experiment was conducted on the King Saud University Saudi Sign Language (KSU-SSL) dataset with no restriction to background, participants' clothes and lighting. This consists of three main phases: input pre-processing, feature learning and feature fusion, and classification. They evaluated the proposed system, conducted experiments in two scenarios, one in Signer-dependent mode: In this scenario, the samples were randomly shuffled and split into two subsets for training and evaluation. and another in Signer-independent mode: In this scenario, the signers were divided into two sets. All the samples performed by the first set of signers were used for training, while all the samples performed by the other set of signers were used for testing. Proposed an efficient deep CNN approach for hand gesture recognition. They proposed an approach employing transfer learning to beat the scarcity of a large labelled hand gesture dataset.

[6] In this paper, they presented a static gesture recognition approach which consists of two stages, a hand pose estimator and a hand pose classifier. The former is used to estimate the hand key points locations, while the latter classifies these predicted locations into different categories. They modified the FGMM (fuzzy Gaussian mixture model) as a classifier

to reject unknown gesture categories and classify the gesture patterns well.

[7] They have proposed two network architectures; the first architecture called Model I has two types of the neural network. The second network architecture, i.e., Model II has a single CNN fed from grayscale or depth data. It has two parallel convolutional neural networks, which are merged by a merge layer, and a RNN with a long short-term memory. The accuracy of the latest model is up to 93%. Pre-processing stage to overcome the variation of video lengths. Deep learning stage to classify, label the frames and recognize the gestures.

[8] To deal with dynamic processes which use a recurrent neural network. They proposed an automatic sampling method which is very useful for a sign language word recognition system. They experimented with a learning data rate of 0.01 and momentum of 0.05. Several improvements such as augmented and filtered data work effectively.

III. PROBLEM STATEMENT

People with a hearing impairment are usually deprived of general communication, as they find it difficult at times to interact with people with their gestures, as only a very few of those are recognized by most people. Also, the general public finds it difficult to understand sign language used by most of the specially abled people. To make this communication effective between specially- abled and general public there is a need to develop a system where both can understand each other.

IV. METHODOLOGY

- We start by collecting key points from mediapipe holistic and collect a bunch of data from keypoints i.e., our hands, on our body and on our face and save data in the form of numpy arrays. We can vary the number of sequences according to our need but each sequence will have 30 frames.
- We then build a LSTM model and train with our stored data which helps us to detect action with a number of frames.
- The number of epochs for the model is determined by us, if we increase the number of epochs the accuracy increases but time taken to run the model also increases and overfitting of model can happen, for gesture recognition.
- Once training is done, we can use this model for real time hand gesture detection and simultaneously convert the gesture to speech using OpenCV.

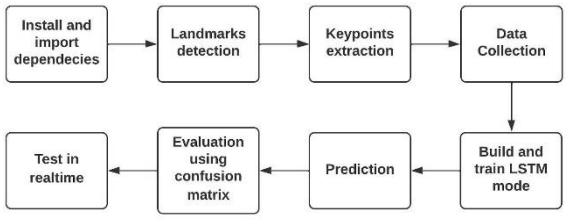

Fig-1: Methodology

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 30, 64)            442112
_________________________________________________________________
lstm_1 (LSTM)                (None, 30, 128)           98816
_________________________________________________________________
lstm_2 (LSTM)                (None, 64)                49408
_________________________________________________________________
dense (Dense)                (None, 64)                4160
_________________________________________________________________
dense_1 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 99
=================================================================
Total params: 596,675
Trainable params: 596,675
Non-trainable params: 0
```

Fig-2: Model Summary

V. OBJECTIVE

- As statistics show that around 80% of specially-abled people are illiterate and the system aims at bridging this gap by converting the sign language known by majority to speech.
- The speech and hearing-impaired people can communicate their message through gestures which can be read out through speech.
- Unimpaired people can use the software to understand sign language and communicate well with specially abled people.
- The project will bridge the gap of difficulty in understanding sign language which was initially present.
- The model will collect the images for gestures through the camera, train the model and check the accuracy for each gesture. Then it will predict the gesture in real-time.

- The project will cover a wide range of gestures for prediction.

VI. TECHNOLOGIES USED

A. Python (3.7.4)

Python is an interpreter, high-level and general-purpose programming language. Python's design system indicates readability of code with its notable use of important whitespace. Its language constructs and object-oriented way aims to help programmers write clear, legitimate code for small and large-scale projects.

B. IDE (Jupyter)

Jupyter Notebook provides us with an easy-to-use, interactive data science environment across many programming languages that not only work as an IDE, but also as a presentation or education tool.

C. 5.3 Numpy (version 1.16.5)

NumPy is a Python library used for working with arrays. It also has functions for working in the dominion of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. NumPy stands for Numerical Python.

D. OpenCV

OpenCV is a vast open-source library for computer vision, machine learning, and image processing. OpenCV supports a large variety of programming languages like Java, C++,
Python, etc. It can process videos and images to identify objects, faces, or even the handwriting of a human. When it is unified with many diverse libraries, such as Numpy which is a highly developed library for numerical operations, then the number of weapons increases in your Arsenal i.e., whatever operations that can be done in Numpy can also be combined with OpenCV. The OpenCV tutorial will help you learn the Image processing from Basics to Advance, like operations on Images, Videos using a huge set of OpenCV programs and projects.

E. Keras

Keras is a high-level neural networks library that is running on the top of TensorFlow, CNTK, and Theano. Using Keras in deep learning allows easy and fast prototyping as well as running seamlessly on CPU and GPU. This framework is written in Python coding language which is easy to debug and allows ease for resilience.

F. TensorFlow

TensorFlow is an end-to-end open-source platform for machine learning. It's a comprehensive and flexible ecosystem of tools, libraries and other resources that provide workflow with high-level APIs. The framework offers many different levels of concepts for you to choose the one you need to build and deploy machine learning models.

G. GTTS

gTTS (Google Text-to-Speech) is a Python library and also a CLI tool to interface with Google Translates text-to-speech API. It writes spoken mp3 data to a file, a file-like object (byte string) for further audio manipulation, or stdout. Or simply pre-generate Google Translate TTS request URLs to feed to an external program.

VII. CONCLUSIONS

Our model, Hand Gesture Recognition, realized that the collection of data base from scratch through video sequences and frames has time constraints and the process is sensitive to gesture changes. The training process would have been achieved faster and we would have an already created database to work off. Some sign language gestures are tough to categorise in our live demo such as "Hello" vs. "I love you" as they differ by a very small change of palm and fingers.
Long Short-Term Memory Network has given an outstanding performance in the detection of sign language hand gestures through video sequences. The model has obtained a categorical accuracy of 80%. We have used 750 epochs and collected 30 video sequences with each sequence containing 30 frames for each gesture. The accuracy may further improve by increasing the database i.e., collecting more videos per gesture.

ACKNOWLEDGEMENT

We wish to thank our parents and associates for their valuable support and encouragement throughout the development of the project work and we would also like to thank our guide Prof. Surendra Sutar for guiding us throughout the project work.

REFERENCES

[1] Yikai Fang, Kongqiao Wang, Jian Cheng and Hanqing Lu, "A REAL-TIME HAND GESTURE RECOGNITION METHOD", *IEEE*, 2007.

[2] N. SWETHA, K. ANURADHA," TEXT-TO-SPEECH CONVERSION", *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 2, no. 6, pp. 269-278, Nov. 2013.

[3] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans and Benjamin Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks", *Springer International Publishing Switzerland*, pp. 572–578, 2015.

[4] Abhishek B, Kanya Krishi, Meghana M, Mohammed Daaniyaal, Anupama H S, "Hand gesture recognition using machine learning algorithms", *Computer Science and Information Technologies*, vol. 1, No. 3, pp. 116-120, Nov. 2020.

[5] Muneer Al-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohamed A. Bencherif, and Mohamed Amine Mekhtiche," Hand Gesture Recognition for Sign Language Using 3DCNN", *IEEE*, vol. 8, April 2020.

[6] Tong Zhang, Huifeng Lin, Zhaojie Ju, Chenguang Yang, "Hand Gesture Recognition in Complex Background Based on Convolutional Pose Machine and Fuzzy Gaussian Mixture Models", Int. J. Fuzzy Syst., 22(4), pp. 1330–1341, Mar 2020.

[7] Falah Obaid, Amin Babadi, Ahmad Yoosofan, "Hand Gesture Recognition in Video Sequences Using Deep Convolutional and Recurrent Neural Networks", *Applied Computer Systems*, vol. 25, no. 1, pp. 57–61, May 2020.

[8] Kouichi Murakami and Hitomi Taguchi Human Interface Laboratory Fujitsu Laboratories LTD. Kawasaki, "Gesture Recognition using Recurrent Neural Networks".