Grenze International Journal of Engineering and Technology, June Issue

Real-Time Gesture Detection for Deaf and Dumb Communication in Video Calls using WebRTC

Pavan Kunchur[1], Krishnaprasad Kulkarni[2], Kishor Balgi[3], Chinmay V Maldar[4], Dhirj Alate[5], Sagar Pujar[6] and Pavan Korlahalli[7]
[1]Associate Professor, Department of Computer Science and Engineering, KLS, Gogte Institute of Technology, Belagavi
Email: pnkunchur@git.edu
[2-7]Department of Computer Science and Engineering, KLS, Gogte Institute of Technology, Belagavi

***Abstract*—In this article we propose to detect the hand and face gestures in a peer to peer video call. First the frames are captured from webcam using openCV, then hand and facial landmarks in video frames are detected using Media Pipe Holistic model. Then the extracted key point values from frames are processed and fed to the trained LSTM model which classifies the gestures into specific texts. The unique feature added is detects accurate text based on highest probability value. Then this is streamed in a peer to peer video call using WebRtc which uses signalling to establish connection among peers using http and SDP protocol.**

***Index Terms*— MediaPipe ,keypoint values, SDP protocol, WebRTC,OpenCV.**

I. INTRODUCTION

Deaf and Dumb individuals have relied on traditional methods like sign language, finger spelling which poses difficulty to understand by normal individuals. Some deaf individuals adopt cochlear implants which is expensive and can range from 5 to 12 INR.

To overcome expensive methods which is not accessible by every dumb and deaf and to enhance the communication we use technology to overcome these problems. By leveraging the technology, particularly within the field of hand gesture recognition and realtime communication, here we propose to reduce the communication gap between deaf, dumb and normal individuals using Realtime Gesture detection in a video call. This impacts by enhancing the quality of life of people with hearing and talking disability and include them in conversation with normal people bridging communication gap between them.

Here we use WebRTC which ensures seamless Real Time Communication between peers. Socket.IO with Node.js and Express handles signalling, which helps in establishment of peer connections. Media pipe is a robust and versatile framework developed by Google Research that enables the implementation of various machine learning-based solutions for multimedia processing tasks. It offers a comprehensive suite of pre-built models and pipelines designed to streamline the development of real-time applications involving tasks such as object detection, pose estimation, and hand tracking.

Towards this the following contributions are made:

1. Performed gesture by user are detected only once with the help of threshold values and probability is set to recognise gesture only once.
2. Integrated Media pipe with WebRTC using ‘recognise sync’ function which optimisesCPU and battery performance of device used.

Grenze ID: 01.GIJET.10.2.520

3. Created a user friendly UI consisting of authorization of users, creation of meeting links, creation of rooms for multiple users using Django and streamlit library.

II. Literature Survey

Dynamic hand and facial gesture recognition in real-time is quite challenging and inaccurate because of limitations of machine learning in vision domain today. Perhaps most significant work is Bogdan Ionescu [2], which offers gesture hand gesture recognition using vision based algorithms but there are two restrictions in this proposal firstly the hand gestures are performed against an approximately uniform background, the distance between the camera and the hand is nearly constant and the gesture is performed in an a priori known region of the space, which is not always possible in real time , and produces wrong results many times .The methodology we use overcomes these limitations and accurately predicts the result. Some of the models proposed, Guillaume Plouffe [3] make accurate hand gesture detection when hands are static but fail to maintain consistency with dynamic gesture detection. We have proposed a technique to reduce error rate by setting probability for gesture in each processed frame and our model only takes last 30 frames of a gesture to predict result to reduce misinterpretation of gesture and enhance accuracy. The model proposed in some articles detect gesture multiple times and introduce additional processing on machine hence increasing processing and power consumption but we propose a technique based on threshold value which avoids multiple detection of same pose or gesture. we use array to store the gesture meanings and display on screen. Vaibhav V [5] has used flex sensor mounted glove for hand gesture recognition, which recognises the hand gestures based on sensor motion and is processed in Raspberry Pi, but the main limitation of this approach is that every deaf and dumb person cannot afford such hardware as they are expensive so we have proposed a software based approach which uses mediapipe holistic model to predict the gesture meaning and they are processed in browser in a videocall using webRTC. The methodology proposed by Bogdan Ionescu [2] is significant the drawback of this proposal is people cannot communicate from a distant place. Our proposed project uses WebRTC for real-time peer to peer connection In a videocall and a normal person can communicate with deaf and mute person over video call. Our proposed work do not use server eliminating privacy issues and establishment of peer to peer connection. This is cost effective as we do not maintain a dedicated server for processing video frames and applying machine learning on them everything is done on edge devices or on peer machine.

III. Proposed Methodology

Architecture

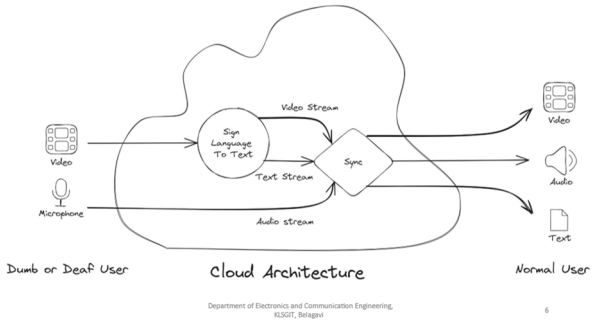

Fig.1 Cloud Architecture

The Entire Project is divided into following sections:
1.Landmark Detection using Media pipe Holistic and computer vision.
2. model building – python LSTM deep learning model.
3.Video Calls Using WebRTC.
4.Integration of Media pipe with WebRTC.

A. LandMark Detection using Media pipe

Media Pipe is open source framework developed by Google which contains tools for building Machine Learning pipelines for multimedia processing tasks. Here we detect hand and face landmarks using hand and face mesh model of solutions module in mediapipe library. After that we set the parameters for the created instance of face

mesh detector and hand pose detector like static_image_mode, min_detection_confidence. Here we read the input image in BGR format and convert into RGB format which is compatible for MediaPipe. After that we use Hand pose detector to detect landmarks from obtained image using process api. Next we iterate through the detected landmark and draw them on image with help of Mediapipe library.After that we extract the keypoint values which contain the cordinates of hand and face landmarks, further we extract the checkpoints with the help of Mediapipe Holistic model and save it into a result variable and we tune the result variable which is compatible with model input feed (Here we flatten result variable and store into numpy array).

B. Model Building – Python LSTM Deep Learning Model

In this Section we have choosen LSTM (Long Short Term Memory) model which is a type of recurrent nural network as we propose to process sequential images(frames). We have used LSTM layers to calssify(multi-class calssification) the squential data (last 30 frames of video capture).We use Tensorflow and keras libraries for creating deep leearning model where we initialize sequential model and we stack layers sequentially ,and use ReLU activation function . The model built contains three sequential layers and two dense layers and one output layer where we use Softmax activation function for multi-class classification. Further we compile the model with optimization algorithm called 'adam' and "categorical_crossentropy" as loss function. Next model is trained with 2000 epoches,taining data point consists of fixed size vector of features, the model contains a call back function called 'tb_callback' which integrates with TensorBoard which is visualization tool by TensorFlow and helps to visualize loss and accuracy metrics .The model iteratively updates the weights and biases and optimizes them based on loss function and optimization alogrithm(adam).whole traning process will be logged and displayed on Tensorboard.At last we predict the model with test data and able to classify each gesture for appropriate meanings(lables for each gesture).The unique feature which we have added is setting the threshold and probablity value for different gestures and the model classifies the input into a label(meaning) which is having more probablity based on last 30 frames of a video sequence, this boosts the accuracy in realtime .

C. Video Calls using WebRTC

Here we propose to create a communication between peers using WebRTC, we use signalling using Socket.IO which helps to exchange messages between client and server using websockets.
The steps involved are:

i. The client who wants to establish connection initates a websocket connection with signalling server using http requests.
ii. The Server acknowledges the request and establishes a WebSocket connection with the client.
iii. The client uses Socket.IO's messaging system to exchange signaling messages with the server and exchange of SDP object which contain session information and finally establish peer to peer connection.

Here signalling server acts as intermediate server which helps to exchange negotiation message between clients like Session Description exchange, ICE candidate exchange which contain IP address and ports of clients in order to establish direct connection between peers. The clients will not be knowing their public IP address so client makes request to STUN server and obtains their public IP address and with public IP they establish peer to peer connection.
The WebRTC components we have used in project are 'MediaStream' to access client's camera and Microphone, RTCPeerConnection for RealTime connection between peers, RTCDataChannel for data transfer between peers. we have used Node.js and express.js for backend of our application.

IV. Integration of Mediapipe with WebRTC

In this section We have integrated the capabilities of mediapipe into peer to peer realtime videocall. Firstly we have captured the video stream with help if getUserMedia() API in WebRTC.Next we have passed the frames of video stream to LSTM model built over Mediapipe Holistic and tensorflow library.Here we have used a event handler to continuously process each frame of video stream in realtime by calling RecogniseAsync function available in WebRTC. This approach gives efficient results without overloading cpu.
RecogniseAsync function operates independently from main execution thread and works asynchronously producing effective results.

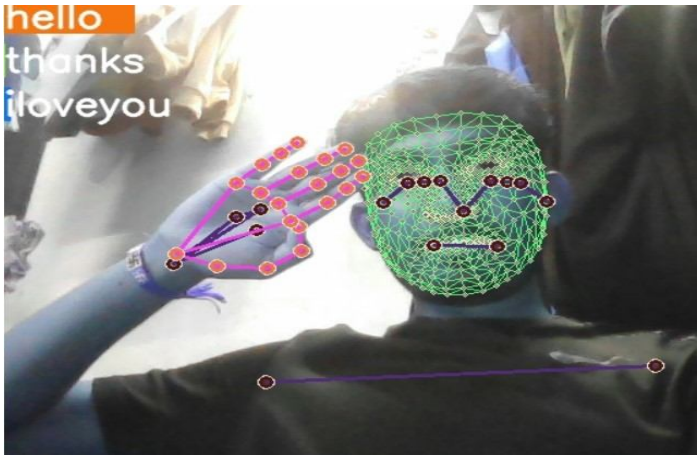

Fig 2. Result Page

V. Results and Discussions

The LSTM model built Accurately predicts Indian Sign Language gestures to text using Mediapipe library. As Mediapipe is cross platform library our project seamlessly runs on all operating systems. The project is sensitive to network bandwidth and accuracy of gesture recognition in video call depends on processing capability of machine used by user.

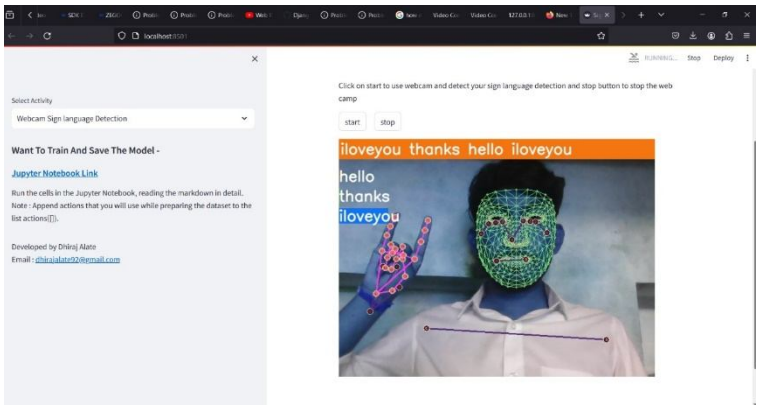

Fig 3. Showing Gesture

Fig 4. Showing Gesture and output

Fig 5. Video chat

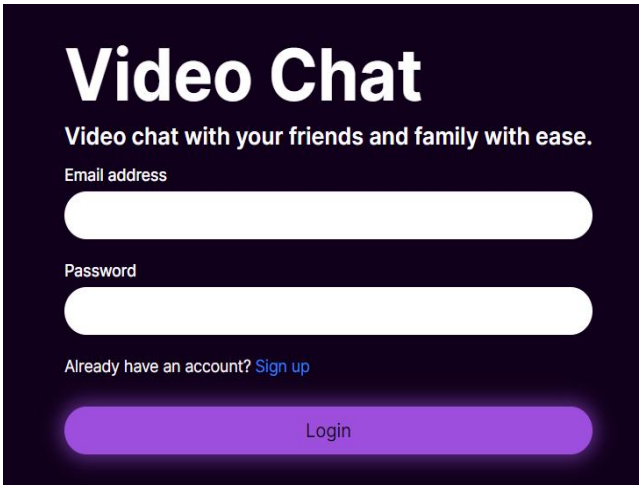

Fig 6. Video Chat Login page

VI. CONCLUSION

In this project we have presented a hand and face gesture detection system over video call without server. We have leveraged the open source project and library of Google called MediaPipe and tensorflow for building LSTM model, further we have used webRTC for peer to peer video communication. This project bridges thecommunication gap between deaf, mute and normal person. The deaf and dumb people can easily convey their thoughts with a normal person from distant place over a video call. Our project efficiently converts hand and facial gestures to text which can be read by normal human which adds value to our society. Our project efficiently detects static and dynamic gesture and efficiently converts into text in realtime.

In Future we can add features which converts text to speech in different languages taking this project to global scale. Further we can use GPUs for decreasing latency in gesture detection which enhances user experience. We have added only Indian sign language detection, in future we can add more sign languages supporting communication dumb and deaf in a global scale.

REFERENCES

[1] Human-machine interactions based on hand gesture recognition using deep learning methods International Journal of Electrical and Computer Engineering (IJECE) Vol. 14, No. 1, February 2024, pp. 741~748 ISSN: 2088-8708, DOI: 10.11591/ijece. v14i1.pp741-748

[2] Dynamic Hand Gesture Recognition Using the Skeleton of the Hand EURASIP Journal on Applied Signal Processing 2005:13, 2101–2109 c 2005 Hindawi Publishing Corporation

[3] Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping Guillaume Plouffe and Ana-Maria Cretu, Member, IEE

[4] Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training Mahdi Abavisani, Hamid Reza Vaezi Joze, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1165-1174

[5] Communication Aid for Deaf and Dumb People. Vaibhav V. Shelke1, Vitej V. Khaire2, Pratibha E. Kadlag3, Dr. K.T.V. Reddy4, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 11 | Nov 2019.

[6] J. Rekha, J. Bhattacharya and S. Majumder, "Shape texture and local movement hand gesture features for indian sign language recognition", *Trendz in Information Sciences and Computing (TISC) 2011 3rd International Conference on.*, 2011.
[7] Zhou Ren et al., "Robust part-based hand gesture recognition using kinect sensor", *IEEE transactions on multimedia*, vol. 15, no. 5, pp. 1110-1120, 2013.
[8] Rung-Huei Liang and Ming Ouhyoung, "A real-time continuous gesture recognition system for sign language", *Automatic Face and Gesture Recognition 1998. Proceedings. Third IEEE International Conference on.*, 1998
[9] Helen Cooper, Brian Holt and Richard Bowden, "Sign language recognition." in Visual Analysis of Humans., London:Springer, pp. 539-562, 2011.