Sign Language Recognition Using Mediapipe Framework with Python

Adarsh Vishwakarma[1] Niraj Yadav[2] Prajnay Yadav[3] Vaibhav Singh[4]
[1,2,3,4]Student
[1,2,3,4]Department of Electronics and Telecommunication Engineering
[1,2,3,4]Shree L.R. Tiwari College of Engineering, Thane, Maharashtra, India

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. Sign language is used by deaf and hard hearing people to exchange information between their own community and with other people. Computer recognition of sign language deals from sign gesture acquisition and continues till text/speech generation. Sign gestures can be classified as static and dynamic. However static gesture recognition is simpler than dynamic gesture recognition but both recognition systems are important to the human community. The sign language recognition steps are described in this survey. The data acquisition, data preprocessing and transformation, feature extraction, classification and results obtained are examined. Some future directions for research in this area are also suggested.

Keywords: Mediapipe, Sign language recognition [SLR], KNN, Hand Solution, Computer Interaction with Humans

I. INTRODUCTION

Sign language using hand gestures is developed for deaf community, which can be used as a means of communication between friends and families of deaf people and the deaf people. Sign Language Recognition is one of the most growing fields of research today as well as challenging too. There are many new techniques that have been developed recently in this field. In this project, we are going to develop a system for conversion of sign language to text using OpenCV. It presents a methodology which recognizes the American Sign Language (ASL) and translates into a normal text.

A. Motivation and Background

SLR aims to develop algorithms and methods to correctly identify a sequence of produced signs and to understand their meaning. Many approaches to SLR incorrectly treat the problem as Gesture Recognition (GR). So research has thus far focused on identifying optimal features and classification methods to correctly label a given sign from a set of possible signs. However, sign language is far more than just a collection of well specified gestures.

B. What is Gesture Recognition?

Gesture recognition is a topic in computer science and language technology with the goal of interpreting human gestures via mathematical algorithms. It is a subdiscipline of computer vision. Gestures can originate from any bodily motion or state but commonly originate from the face or hand. Current focuses in the field include emotion recognition from face and hand gesture recognition. Users can use simple gestures to control or interact with devices without physically touching them. Many approaches have been made using cameras and computer vision algorithms to interpret sign language.

C. Sign Language

Sign languages [also known as signed languages] are languages that use the visual-manual modality to convey meaning. Sign languages are expressed through manual articulations in combination with non-manual elements. Sign languages are full-fledged natural languages with their own grammar and lexicon. Sign languages are not universal and they are not mutually intelligible with each other, although there are also striking similarities among sign languages.

Linguists consider both spoken and signed communication to be types of natural language, meaning that both emerged through an abstract, protracted aging process and evolved over time without meticulous planning. Sign language should not be confused with body language, a type of nonverbal communication.

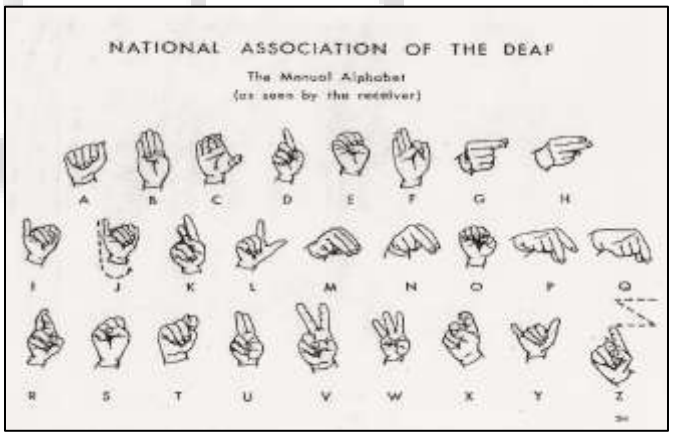

Fig. 1: Sign Language

D. Mediapipe Framework [Hand Solution]

MediaPipe Hands is a high-fidelity hand and finger tracking solution. It employs machine learning (ML) to infer 21 3D landmarks of a hand from just a single frame. Whereas current state-of-the-art approaches rely primarily on powerful desktop environments for inference, our method achieves real-time performance on a mobile phone, and even scales to multiple hands. We hope that providing this hand perception functionality to the wider research and development community will result in an emergence of creative use cases, stimulating new applications and new research avenues.

MediaPipe Hands utilizes an ML pipeline consisting of multiple models working together: A palm detection model that operates on the full image and returns an oriented hand bounding box. A hand landmark model that operates on the cropped image region defined by the palm detector and returns high-fidelity 3D hand keypoints. This strategy is

similar to that employed in our MediaPipe Face Mesh solution, which uses a face detector together with a face landmark model.

E. Objective

The goal of this project was to build a neural network able to classify which letter of the American Sign Language (ASL) alphabet is being signed, given an image of a signing hand. This project is a first step towards building a possible sign language translator, which can take communications in sign language and translate them into written and oral language. Such a translator would greatly lower the barrier for many deaf and mute individuals to be able to better communicate with others in day to day interactions.

F. Summary

Developing sign language application for deaf people can be very important, as they'll be able to communicate easily with even those who don't understand sign language.Our project aims at taking the basic step in bridging the communication gap between normal people and deaf and dumb people usingsign language. The main focus of this work is to create a vision based system to identify Finger spelled letters of ASL. The reason for choosing a system based on vision relates to the fact that it provides a simpler and more intuitive way of communication between a human and a computer. In this report, 36 different categories have been considered: 26 categories for English Alphabets (a to -z) and 10 categories for Numerals (0-9).

We used Mediapipe FrameWork, Google's Mediapipe Solutions has developed a Hand Recognition model and it can detect 21 3D Landmarks of Palm. So we will try to understand it and how we can use this Library with Python so as to achieve our Goal. It achieves an average precision of 95.7% in palm detection. Using a regular cross entropy loss and no decoder gives a baseline of just 86.22%.

II. LITERATURE SURVEY

In recent years, there has been tremendous research on hand sign language recognition. The technology of gesture recognition is divided into two categories:-

A. Vision Based Approach

This approach takes images from the camera as data of gesture. The vision based method mainly concentrates on captured images of gestures and extract the main feature and recognizes it. The color bands were used at the start of the vision based approach. The main disadvantage of this method was the standard color should be used on the fingertips. Then use of bare hands rather than the color bands. This poses a challenging problem as these systems need to be background invariant, lighting insensitive, person and camera independent to achieve real time performance. Moreover, such systems must be optimized to meet the requirements, including accuracy and robustness.

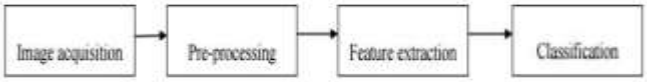

Fig. 2: Block Diagram of vision based recognition system

Vision based analysis is based on the way human beings perceive information about their surroundings, yet it is probably the most difficult to implement in a satisfactory way. Several different approaches have been tested so far.

One is to build a three dimensional model of the human hand. The model is matched to images of the hand by one or more cameras, and parameters corresponding to palm orientation and joint angles are estimated. These parameters are then used to perform gesture classification.

Second one is to capture the image using a camera then extract some features and those features are used as input in a classification algorithm for classification.

B. Sensor Based Approach

This approach collects the data of gestures performed by using different sensors. The data is then analyzed and conclusions are drawn in accordance with the recognition model. In case of hand gesture recognition different types of sensors were used and placed on hand, when the hand performs any gesture, the data is recorded and is then further analyzed. The first sensor used was Data gloves then LED's came into existence. The invention of the first data glove was done in 1977. Sensor based approach damages the natural motion of the hand because of use of external hardware. The major disadvantage is complex gestures cannot be performed using this method

In this system an electro mechanical robot is designed and controlled using hand gestures in real time. The system is designed on a microcontroller platform using Keil and MPLAB tools. Hand gesture recognition is done on the principle of resistance change sensed through the flex sensor. These sensors are integrated in hand gloves from which input to the system is given. The designed system is divided into two sections as transmitter and receiver. The transmitter section will be in hand gloves from which the data is sensed and processed through PIC16F7487 and sent serially to the receiver section. RF technology is used to transmit the data at the receiver section at the frequency of 2.4 GHz. ARM 7 (LPC2148) processor is used to receive the data. Here from the received data, the character is predicted and matched with the closest character from which the character is identified and displayed on LCD. The various case studies are prepared for the designed system and tested in real time. The proposed system can be used for various applications such as in unmanned machines, industries, handicapped personnel etc.

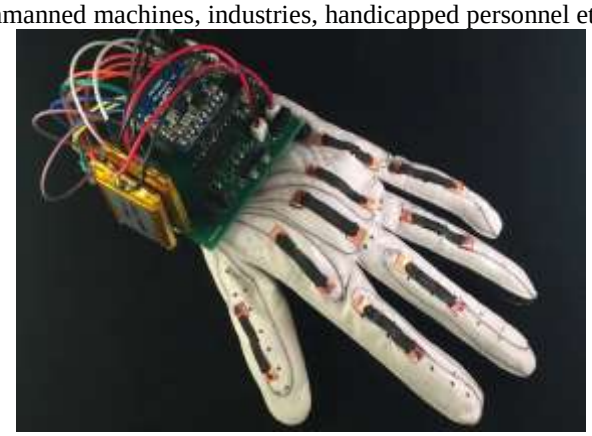

Fig. 2: Gloves with flex sensor mounted on it

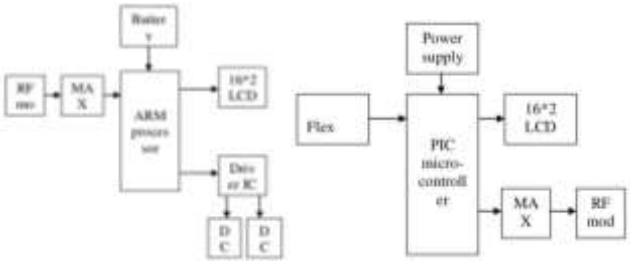

Fig. 6: Transmitter and Receiver Section of Hand Gesture

III. Implementation Methodology

MediaPipe Hands utilizes an ML pipeline consisting of multiple models working together: A palm detection model that operates on the full image and returns an oriented hand bounding box. A hand landmark model that operates on the cropped image region defined by the palm detector and returns high-fidelity 3D hand keypoints.

So to make a Database we have to Capture at most 100 images per Sign and with this model we can get 21 keypoint of the hand. i.e. It's coordinates [x, y, z]. x and y are normalized to [0.0, 1.0] by the image width and height respectively. z represents the landmark depth with the depth at the wrist being the origin, and the smaller the value the closer the landmark is to the camera. After making a Database we can predict the sign with the help of Suitable Model. We will use the KNN algorithm.

A. Hardware & Software Requirement

1) Computer with windows or Linux, Python Installed in it and Libraries.
2) CMOS Sensor (WebCam)
3) Hand Gestures for Recognition

Hardware & Software that we have used in our Project Hand Sign Recognition:

1) Python Installed in Windows Os or Linux Os Machine.
2) CPU - Intel core i5 9th Gen.
3) GPU - Nvidia GTX 1050 Ti.
4) Webcam 720p60
5) Python 3.8.6 and It's IDE like VS, Spyder etc
6) Libraries : OpenCV, Tensorflow, Keras, MediaPipe and lots more basic*
7) KNN (The k-nearest neighbors) from Sklearn library of Python

B. Result

This sign language recognizer is able to recognize alphabets (A-Z) and numbers (0-9). All the signs can be recognized real-time. The current system has only been trained on a very small database so some of them didn't get recognized.

Accuracy using KNN algorithm on Single handed Gesture is 92.30769230769231 %

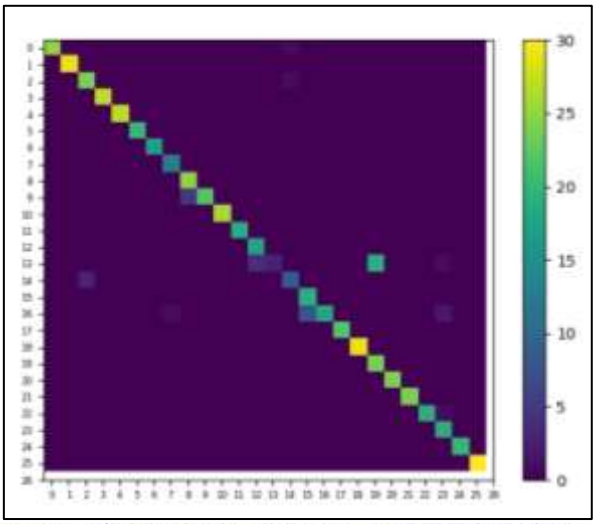

Fig. 4: KNeighbors Classifier Confusion Matrix

Accuracy using Logistic Regression on Single handed Gesture is 86.78929765886287 %

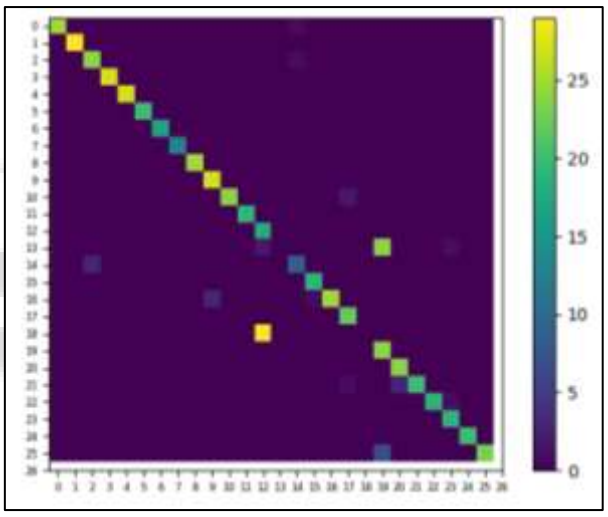

Fig. 5: Logistic Regression Confusion Matrix

Accuracy using DecisionTree Classifier on Single handed Gesture is 92.47491638795987 %

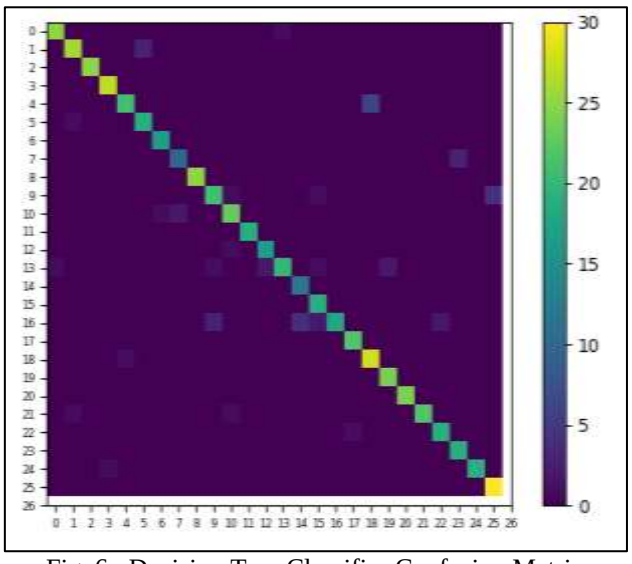

Fig. 6: Decision Tree Classifier Confusion Matrix

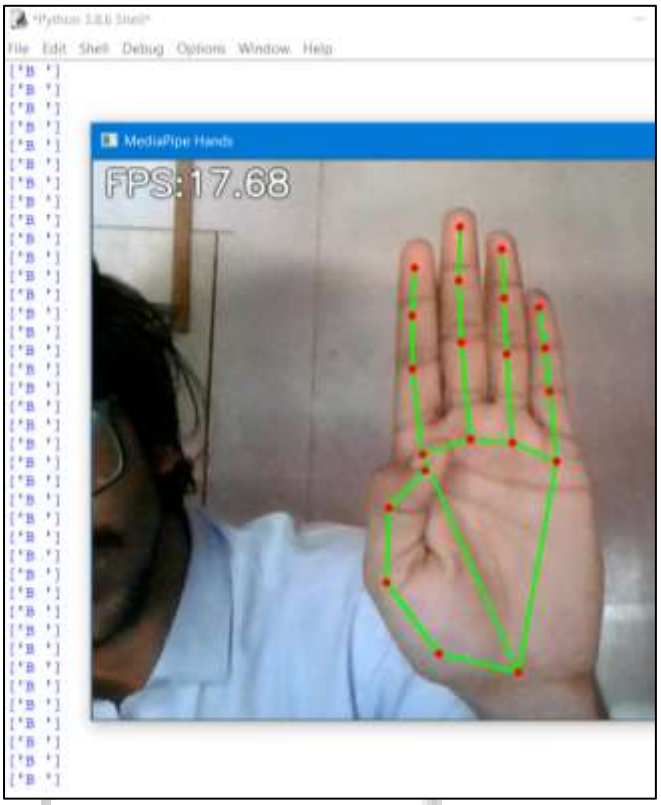

Fig. 7: Letter B

Average Image Processing and Backend Process is about 17 FPS which can be further increase by changing Dataset, Algorithm, and Hardware.

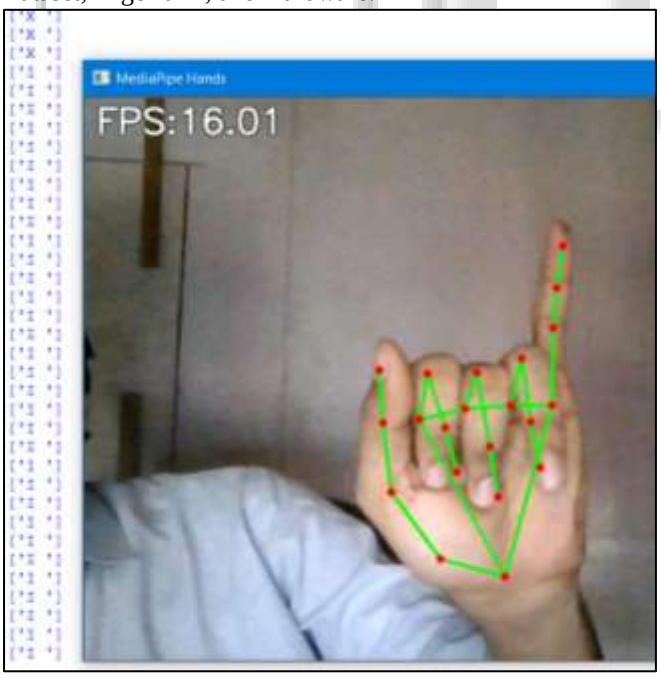

Fig. 8: Letter I

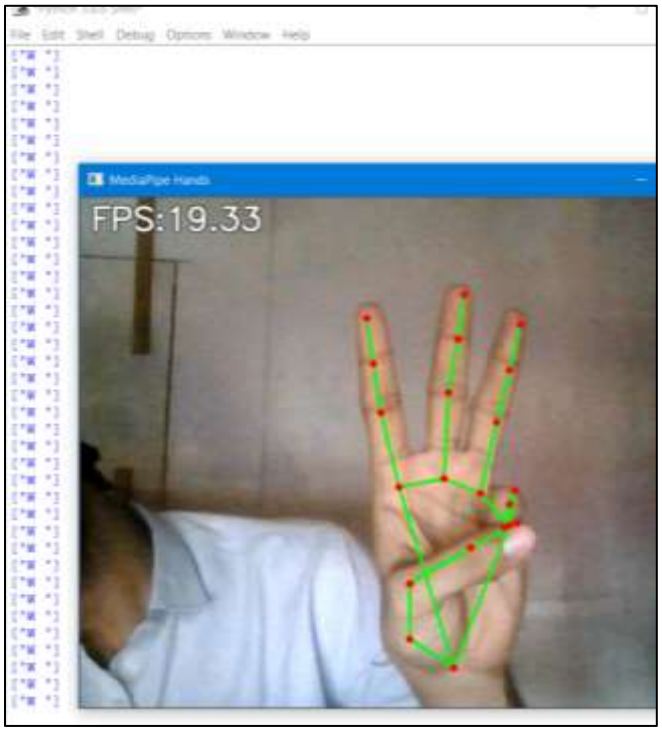

Fig. 9: Letter W

IV. CONCLUSIONS

Sign Language using Mediapipe and recognition through computer vision was a partly successful one with an accuracy rate of average 17 FPS and accuracy on an average is 86 to 91 %. The question of perfection is another quest to deal with in the days to come. The hand gesture detection and recognition were the main topic and problem that were dealt with. Mediapipe achieve an average precision of 95.7% in palm detection. Using a regular cross entropy loss and no decoder gives a baseline of just 86.22%. So the Future Scope of this Research will be Good to develop Human Computer Interaction (HCI) by using more powerful and fast algorithm.

REFERENCES

[1] https://ai.googleblog.com/2019/08/on-device-real-time-hand-tracking-with.html
[2] https://google.github.io/mediapipe/solutions/hands.html
[3] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati, and M. Hadap, "Hand Gesture Recognition for Indian Sign Language", IEEE International Conference on Computer Communication and Informatics (ICCCI), Jan. 10-12, 2012, Coimbatore, India.
[4] Zhi-hua Chen, Jung-Tae Kim, Jianning Liang, Jing Zhang,, and Yu-Bo Yuan, "Real-Time Hand Gesture Recognition Using Finger Segmentation" IEEE International Conference on department of Computer Science and Engineering, East China University of Science and Technology, Published 25 June 2014 Shanghai 200237, China
[5] Umang Patel & Aarti G. Ambekar "Moment based sign language recognition for Indian Languages "2017 Third International Conference on Computing, Communication, Control and Automation (ICCUBEA)
[6] https://en.wikipedia.org/wiki/Sign_language

