International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Gesture Tool – Aiding Disabled Via AI

Abhishek Kumar Saxena[1], Krish Shah[2], Mayank Soni[3], Shravani Vidhate[4]

[1-4]Student, Dept. of Computer Science and Engineering, MIT School of Engineering, MIT ADT University, Pune Maharashtra, India

---

Abstract - *Communication is one of the most important things when it comes to living in a society, well, average people do not suffer much because they are able to interact properly with those around them. Unfortunately, there are a few people who are disabled, some who can neither speak nor hear. Most of us also wonder how they interact as we think about how deaf individuals live. Although sign language makes it easier for them to communicate with each other it also establishes a barrier between deaf and dumb individuals and ordinary individuals. Since most hearing people do not know how to "speak" the sign language and have little patience to practice, interacting with them is generally inconvenient for deaf people. The purpose of this project is to build a real-time hand gesture recognition system using concepts of Deep Learning and NLP that recognizes hand gestures and translates gestures into speech, as well as text message, and vice versa.*

***Key Words*: Communication, Deep Learning, NLP, Gesture, Recognition, Sign language**

1.INTRODUCTION

Communication between humans is an action that can be seen as more than just the transfer of knowledge to one another. We, as people, take more than just the truth to communicate. The facial expressions, the language of the body, the way you speak, the way you react, the way you go with the language, the thought, the meaning, the choice of words, all constitute to how you communicate. Some people are either born with defects that render them unable to hear or speak, or an impairment caused by an accident that makes them unable to talk or hear. They do not understand what their own speech sounds like or the meaning of words cannot be understood. The gap between normal human beings and deaf and dumb is wide and ever-increasing day by day. They can't understand what other people say, or they can't articulate themselves in a way that is understandable, even though they understand it. Over the years, these people have been creative enough to come up with several ways to express themselves, and have introduced sign language to help them learn to communicate. This was a very significant step in improving communication skills for deaf and dumb people. Sign language provides the deaf and dumb people with the best communication medium to communicate with people. Sign language is difficult to understand for people who are not usually well aware of it. Gesture is a non-verbal communication that requires hand gestures that are much preferable and easier. The system is based on the hand gesture method to computer vision.

2. LITERATURE SURVEY

[1] In this paper, they have suggested a system for understanding hand movements that would not only act as a way of communicating between deaf and dumb and mute individuals, but also as an educator. Being very common among these individuals, hand gestures serve as a good means of communication. The hand is therefore considered as an input for a system that will either display the corresponding result in the form of text or voice, or both. The framework for gesture recognition mainly consists of image acquisition, segmentation followed by morphological erosion and gesture recognition feature extraction.

[2] A proposed system that is simple to implement without a complicated calculation of features. With 90% precision, they measured the high gesture recognition rate. Their approach projects hand gestures into speech and vice versa, with the use of image acquisition, image processing, extraction of features and a few more principles of artificial intelligence and data mining.

[3] It implies a vision-based technique in which continuous sequences of complex hand gesture frames are taken. The four primary steps for identifying complex hand gestures are pre-processing, segmentation, feature extraction and classification. The technique of skin colour filtering for segmentation was used in their method. The technique of Eigen vectors and Eigen values has been used for feature extraction. For more detailed feature identification, they have also used Histogram matching. Word sign prediction using one or both hands, dynamic hand gesture words dataset working with Indian Sign language and dual-way communication has also been proposed in this framework.

[4] The framework extracts gestures from real-time video in the Indian Sign Language (ISL) and maps them with a human clear and understandable voice. They have used the video-to-speech processing technique that will involve video frame creation, region of interest (ROI) discovery and information domain image mapping using the Correlational-based approach and then related audio generation using Google Text-to-Speech (TTS) API. By translating speech to text using the Google Speech-to-Text (STT) API, the tongue is mapped with equivalent Indian signing gestures, further mapping the text to the related animated gestures from the database. Real-time video (ISL) translation to human natural language speech equivalent.

[5] This translator uses the camera to input the gesture, translates the result and spells it out. These were implemented using the deep learning (CNN model) and OpenCV libraries supported by the Raspberry Pi 3 model. To retrieve the individual frames, the camera module captures the video and sends it to the controller. This module measures 25 mm x 20 mm x 9 mm and weighs around 3 g, making it suitable for portable devices. This module has a sensor with a fixed focal length lens. This camera captures the image and transfers it to the Raspberry Pi camera port. The picture is interpreted by the Raspberry pi and determines the symbol that is further translated into expression. The surrounding people will understand and respond to the sign language used by the person (deaf-mute) carrying this prototype, as the voice is rendered audible using a speaker. 80% of the dataset is trained on the CNN model of the proposed framework and the remaining 20% of the data is used for cross-validation. The model is trained for 20 epochs and 99% is observed to be the accuracy of the trained model.

[6] Following the Sign Language Standard, the suggested solution addresses the meaning of sentences. For the sign recognition system, it requires intensive preparation such that the words recognized can be organized as a coherent sentence. The measures like POS tagging, Parsing, Disambiguation, Sentence boundary, Text to Speech are used by the NL framework (TTS). For grammatical analysis of a given sentence, the parse tree is used. An average of 80% accuracy is given by this method.

[7] Using a Chinese sign language recognition scheme in real time. A Chinese sign language dataset was developed, and a video stream was captured using an RGB camera. To extract feature vectors, they used a 3D-CNN method. For better comprehensive hand and head detection efficiency, they used YOLO. This approach used 3 distinct methods, but 3D-CNN+RGB+Optical flow provided 90.1% of the best accuracy.

[8] To record 3D hand and finger movements, a sensor-based motion tracking device is used. They proposed a completely unique angular velocity technique, which is directly applied to real-time 3D motion data streamed by the sensor-based system to detect and recognize hand movements. The technique is capable of real-time identification of both static and dynamic movements. With two interactive applications that need gesture feedback to communicate with the virtual world, they determine the precision of popularity and execution efficiency. Their experimental results indicate high accuracy of recognition, high performance of execution, and high usability levels.

[9] Detection of fingertips and real-time hand motion recognition using an RGB-D camera and a 3D convolution neural network (3DCNN). The device can extract fingertip positions precisely and robustly and recognize movements in real-time. By evaluating hand gesture recognition across a variety of gestures, they demonstrated the accuracy and robustness of the interface. In addition, they developed a software program control mechanism to point out the possibility of using hand gesture recognition. The experimental findings showed that our device features a high degree of recognition of hand gesture accuracy.

[10] Portable technology and Arduino Circuit Boards are designed to provide various people with one or more of the above-mentioned disabilities with a means of communication. A type of wearable technology is the scheme they have used. The input can be text for a blind person using the Sensor Glove or Braille script keyboard as per his/her wishes and specifications, and the audio and Braille pattern can be the output. Thus, the device is set so that it can take input and give output as per the requirements of the user. The message to be sent by the user is taken as an input to the gadget. The computer is therefore set so that it can take input and provide output according to the user's requirements. As an input to the gadget, the message to be sent by the user is taken. Text, gesture, or Braille language can be the input. The gadget has a text input Sensor Glove and a Braille Language Converter for taking and translating the Braille Language input into text. If the message to be transmitted by the sender is in the form that is appropriate and understandable by the recipient and the communication is a direct type of Communication, then the message is transferred directly to the receiver.

3. ALGORITHMS AND FEATURES

We have used Google Speech API to convert the audio input to text. It enables you to convert short-form or long-form audio to text to offer unmatched accuracy. You can enable voice searches (such as "How is the weather outside"), command use cases (such as "Start playing music"), and transcribe audio from calls. In addition, the API supports 120 languages and dialects throughout the world. In this way, developers can enhance their app's functionality and build intelligent systems that recognize speech.

To make this project more efficient, we have used Mediapipe along with other algorithms. Mediapipe is a framework for building multimodal (e.g., video, audio, any time series data), cross-platform (i.e., Android, iOS, web, edge devices) applied machine learning pipelines. From Mediapipe, we have used two models known as the Palm Detection model that detects the palm of the hand using the full image and returns an oriented bounding box of the hand. and the Hand Landmarks Model that returns a 3D keypoint for the hand using the cropped region defined by the palm detector.

Palm Detection Model: A single-shot detector model was built to detect initial hand locations. Rather than using a hand detector, we trained a palm detector, since estimating bounding boxes of rigid objects such as fingers and palms is more straightforward than detecting hands with articulated fingers. In ML terminology, a palm can be modelled by using square bounding boxes (anchors), which ignore other aspect ratios and reduce the number of anchors between

three and five. To avoid the high size variance of the objects, we use an encoder-decoder feature extractor to extract the attributes of large scenes (similar to RetinaNet). Lastly, we minimize the focal loss during training to support a large number of anchors.

A regular cross entropy loss plus no decoder yields an average precision of 95.7% in palm detection, while the above two methods generate an average precision of 86.22%.

Hand Landmark Model: As a result of palm detection for the entire image, our hand landmark model performs regression to automatically determine 21 3D keypoints for hand placement within the detected hand regions, i.e., direct coordinate prediction.

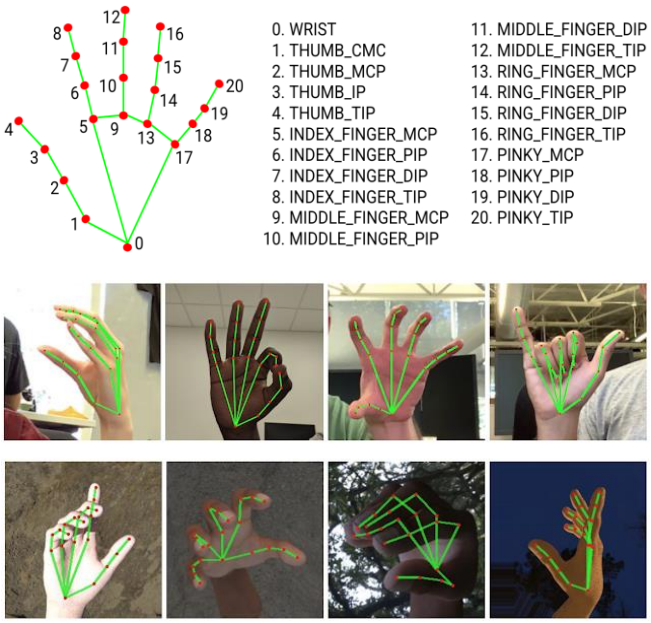


Fig 1. With keypoints annotation, aligned hand landmarks are sent to the tracking network.

For Large Vocabulary Speech Recognition, Generation of Features, and Sentence Constructions, we used the following two algorithms: Single Shot Detector (SSD), Long Short-Term Memory Based Recurrent Neural Network Architecture (LSTM RNN).

Single Shot Detector (SSD)

In SSD, each feature map location discretizes the output space of bounding boxes into a set of default boxes at various aspect ratios and scales. The network generates scores for the presence of each item type in each default box at prediction time, and then adjusts the box to better match the object shape.

Across the entire image, we have several default boxes of varying sizes as well as aspect ratios. 8732 boxes are used by SSD. This assists in the estimation of the default bounding box that most strongly matches the ground truth bounding box comprising objects. During training, the default boxes are matched to the ground truth boxes in terms of aspect ratio, location, and scale. The boxes exhibiting the greatest overlap with the ground truth bounding boxes are chosen. Between anticipated boxes and ground truth, the IoU (intersection over union) should be bigger than 0.5. Finally, we choose the predicted box having the greatest overlap with the ground truth.

Long Short-Term Memory Based Recurrent Neural Network Architecture (LSTM RNN)

The Long Short-Term Memory (LSTM) architecture is a recurrent neural network (RNN) architecture that was implemented to address the vanishing and exploding gradient issues that plague conventional RNNs. RNNs have cyclic connections, unlike feedforward neural networks, which makes them useful for simulating sequences.

An RNN with LSTM units can be instructed in a supervised on a set of training sequences, using an optimization algorithm including such gradient descent combined with backpropagation through time to compute the gradients needed during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network).

4. PROPOSED STUDY

The core principle behind addressing such a scenario is to use Deep learning to do multi-gesture recognition while designing a cross-platform web application. Building a web application that enables computer vision and natural language processing to assist the deaf community communicate more effectively by removing the barrier of losing a sense through gesture detection.

In order to address the second impairment, it uses Google API to transform voice translation and speech to text over the device's screen. The model is put on the cloud for real-time gesture interpretation in order to construct words from a collection of gestures and vice versa.

The processing of three modules of the project are mentioned below:

Module 1: Gesture to audio/text

1. Input: Takes input from a webcam via a live video stream.

2. Image Acquisition: Frames from a video stream were recorded for image acquisition.
3. Image Preprocessing: The query image will be in RGB format. Grayscale images will be created from these RGB colour photographs.
4. Feature Extraction: The RGB picture is used to identify key points. All key points have their Euclidean distance evaluated. These measurements are then standardised and used as features.
5. Output (Gesture to Speech/Text transition): Uses video processing techniques such as video frame creation, region of interest (ROI), and knowledge-domain mapping of pictures, as well as the Google Text-to-Speech (TTS) API to generate relevant audio/text.

Module 2: Text/audio to gesture

Speech or text is used to provide input. If the input is in speech, the system uses the Google Speech API to transform it to text. The text will be preprocessed, and the important keywords from the sentences will be recognised. The backend will execute the translation system, and data will be retrieved. It will conduct a quick video with the animation for each of the detected keywords and display it.

Module 3: Self Training the modules

We enter which gesture we are aiming to make during our initial data collection phase, then select the range of samples we require, and ultimately begin collecting data.

We then proceed to the training phase, where we define the train-to-test ratio, the number of layers, and the number of neurons in each layer, and begin training.

After training, we move on to the testing and evaluation stage, when we switch on our camera and display model gestures for it to recognise.

5. EVALUATION

1. **Data set and data features:** For this project, we have built our own dataset. We captured a large amount of training snapshots for each hand gesture, which are the fist, index, palm, and little finger gestures, for different people, scales, and rotations, and under varying lighting circumstances. The hand gesture will be represented by keypoints extracted from training images using the Euclidean algorithm.

   In the Euclidean algorithm, Euclidean distance is the real distance between two points in 2D and 3D space. It is also the distance between two points in n-dimensional space. Euclidean distance is defined as:

$$d=\sqrt{\sum - 2\,d\,(yixi)^2}, \text{ i = 1, 2... n.}$$

   An image can be viewed as a matrix of pixel values for a hand gesture image, or can be extended to be viewed as a vector, such as a N x N pixel image can be viewed as a vector of length N^2, with each pixel indicating a point in the N^2 dimensional space. When d is less than a specific threshold, the Euclidean distance between two images is defined as $d = ||\,\Omega - \Omega k\,\,||^2$, indicating that this image has been classified as the k-th kind hand gesture.

2. **Confusion Matrix:** A confusion matrix, also known as an error matrix, is a precise table arrangement that permits evaluation of the performance of an algorithm, typically a supervised learning one, in the field of machine learning and specifically the problem of statistical classification (in unsupervised learning it is usually called a matching matrix). The cases in an actual class are defined by each row of the matrix, whereas the instances in a predicted class are indicated by each column, or vice versa.

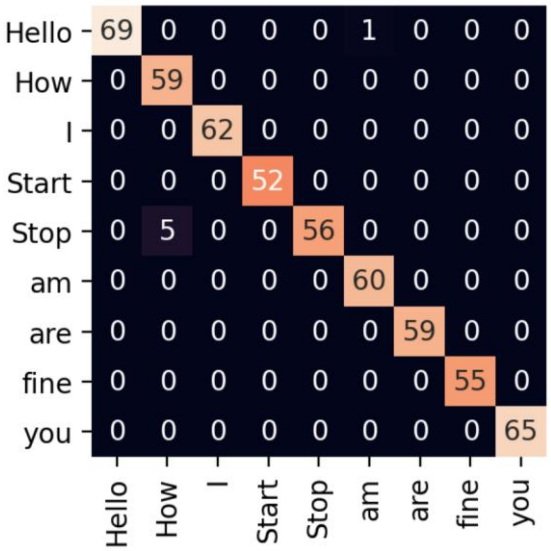


Fig 2. Confusion Matrix

The confusion matrix in Fig. 2 is used to demonstrate the CNN model's performance. The true label is seen in each row of the confusion matrix, while the predicted label is seen in each column. For example, this confusion matrix shows that out of 61 images of "Stop" used for testing the model, only 56 are correctly predicted as "Stop" and 5 are predicted incorrectly.

3. **Experimental Results:** We obtained an accuracy of 99.81%, which can be computed as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The web application has successfully been able to recognise the gestures and display phrases in text, audio as well as in animated clips with 99.81 % accuracy. But sometimes, accuracy tends to be an incorrect metric for such algorithms. Therefore, we use the other evaluation metrics which are Precision and Recall.

Precision is the ratio of true positive detection to total number of positive detection, which can be computed as:

$$Precision = \frac{TP}{TP + FP}$$

Recall is the ratio of true positive detection to total number of detection, which includes both true positive and false negative detection, which is computed as:

$$Recall = \frac{TP}{TP + FN}$$

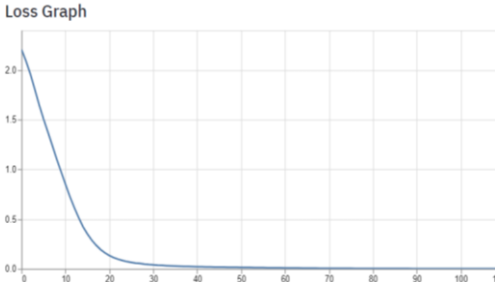
precision recall f1-score support

0	1.00	0.99	0.99	70
1	0.92	1.00	0.96	59
2	1.00	1.00	1.00	62
3	1.00	1.00	1.00	52
4	1.00	0.92	0.96	61
5	0.98	1.00	0.99	60
6	1.00	1.00	1.00	59
7	1.00	1.00	1.00	55
8	1.00	1.00	1.00	65
accuracy			0.99	543

4. **Loss Graph;** The following is the loss graph for model training. The optimizer used for minimizing the loss function is Adam. The loss function used for this model is Log Loss and it is computed as:

$$Logloss_i = -[y_i \, ln \, p_i + (1 - y_i) \, ln(1 - p_i)]$$

where i is the given observation/record, y is the actual/true value, p is the prediction probability, and ln refers to the natural logarithm (logarithmic value using base of e) of a number.



6. CONCLUSIONS

A real-time sign language framework based on RGB video streams and a dataset of sign languages is created in this project, which provides the basis for a sign language recognition system. For hand gesture detection, which has better comprehensive efficiency, a combination of mediapipe SSD hand-keypoint model with a neural network classifier. The framework tries to lessen communication gap, the difference between deaf people and the normal world, since dual contact is encouraged.

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude towards my teacher Prof. Reena Pagare for providing us with the golden opportunity to make this wonderful project on the topic 'Gesture Tool - Aiding disabled via AI', which also helped us in doing a lot of research and learning about new things.

We are highly indebted to MIT School of Engineering for their guidance and support as well as providing the necessary information and resources regarding the project.

REFERENCES

[1] AAWAAZ: A Communication System for Deaf and Dumb: A. Sood and A. Mishra, "AAWAAZ: A communication system for deaf and dumb," 2016 5th International Conference on Reliability, Info com Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 620-624, DOI: 10.1109/ICRITO.2016.7785029.

[2] Real Time Two Way Communication Approach for Hearing Impaired and Dumb Person Based on Image Processing: S. S. Shinde, R. M. Autee and V. K. Bhosale, "Real time two-way communication approach for hearing impaired and dumb per son based on image processing," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai,2016, pp.1-5, DOI:10.1109/ICCIC.2016.7919572.

[3] Development of Full Duplex Intelligent Communication System for Deaf and Dumb People: S. Rathi and U. Gawande, "Development of full duplex intelligent communication system for deaf and dumb people," 2017 7th International Conference on Cloud Computing, Data Science & Engineering- Confluence, Noida, 2017, pp. 733-738, doi:

10.1109/CONFLUENCE.2017.7943247.

[4] Two Way Communicator Between Deaf and Dumb People and Normal People: P. G. Ahire, K. B. Tilekar, T. A. Jawake and P. B. Warale, "Two Way Communicator between Deaf and Dumb People and Normal People," 2015 International Conference on Computing Communication Control and Automation, Pune, 2015, pp. 641-644, doi: 10.1109/ICCUBEA.2015.131.

[5] Assistive SIGN LANGUAGE Converter for DEAF AND DUMB: L. Boppana, R. Ahamed, H. Rane and R. K. Kodali, "Assistive Sign Language Converter for Deaf and Dumb," 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (Green Com) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (Smart Data), Atlanta, GA, USA, 2019, pp.302-307,DOI:10.1109/iThings/GreenCom/CPSCom/SmartData 019.00071.

[6] Multimodal Interface for Deaf and Dumb Communication: S. S. Wazalwar and U. Shrawankar, "Multimodal Interface for Deaf and Dumb Communication," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2019, pp. 418-422.

[7] Real-Time Sign Language Recognition Based on Video Stream: K. Zhao, K. Zhang, Y. Zhai, D. Wang and J. Su, "Real-Time Sign Language Recognition Based on Video Stream," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 7469-7474, doi: 10.23919/CCC50068.2020.9188508.

[8] Design and evaluation of a hand gesture recognition approach for real-time interactions (2020): O. Kopuklu, A. Gunduz, N. Kose and G. Rigoll, "Real time Hand Gesture Detection and Classification Using Convolutional Neural Networks," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019, pp. 1-8, doi: 10.1109/FG.2019.8756576.

[9] Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network (2020): Tran, Dinh-Son; Ho, Ngoc Huynh; Yang, Hyung-Jeong; Baek, Eu-Tteum; Kim, Soo-Hyung; Lee, Gueesang. 2020. "Real Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network." Appl. Sci. 10, no. 2: 722.

[10] A NOVEL APPROACH AS AN AID FOR BLIND, DEAF AND DUMB PEOPLE: B. Rajapandian.