Using Hand Gesture Cognizance and Voice Instruction for Human Computer Interaction

M Nithya[1]
nithya.cse@sairam.edu.in
Computer Science and Engineering
Sri Sairam Engineering College

Eyuwankg S Swankg[2]
messikarthik13@gmail.com
Computer Science and Engineering
Sri Sairam Engineering College

M Kaviyarasan[3]
e8cs119@sairamtap.edu.in
Computer Science and Engineering
Sri Sairam Engineering College

Abstract.

The human-computer interaction is one of the arduous problems for disabled people and usually, the way to interact with the computers is by using a physical mouse and a keyboard, But this poses a hindrance for disabled people moreover it's not a great experience for the user who is in the architectural and design field. There are some tracking systems available to use for hand recognition, but they are slow and expensive to use. The development of a hand tracking and gesture recognition system with no marker state is implemented with novel technology. The proposed system is easy to use and efficient which allows it continuous tracking of hands even with any type of background. This system removes motion blur and can detect various gestures and different actions for that gestures using computer vision techniques. In order to provide the interactive platform, the detected gesture has a specific set of actions that are executed by the system in addition to that voice commands have been used to give input action to the system. A voice recognition module is implemented to get voice command input from the user and execute the action. This provides an interactive HCI and intuitive System.

Keywords. Human Computer Interaction, Machine Learning, Voice recognition, Gesture recognition, Natural Language Processing.

1. INTRODUCTION

Human-computer interaction is one of the arduous problems that stayed in a traditional way which didn't give any great experience to the user and has some restrictions in terms of usage manner. The way of interacting with the computer is by using regular keyboards and mouse which need to be carried around and in laptops, it adds up as additional space. There are some hand tracking systems available to use, but they are slow and highly-priced. This paper describes the development of hand tracking system with no marker along with gesture recognition feature using very low price hardware components. Hand gestures are a type of nonverbal communication. It also has a significant impact on Systems for human-computer interaction (HCI). As a result, there is a huge demand for automatic handgesture recognition system. Since the rise of machine learning and computer vision systems has increased, it provides a better solution to these problems. From video games to control systems, HCI offers a broad range of applications. Because of their temporal dependence, Hand gestures, like other time-varying signals, can't be compared directly in Dimensional space. This relationship reveals crucial distinguishing characteristics. It's hard to retrieve meaningful hand-engineered features for hand motions because of temporal misalignment and huge irrelevant portions in every frame. To overcome this, ready-to-use media pipe solutions can be used and on top of that existing solution, a custom model can be built to get the accuracy required for the applications. The required hand gestures are takenfrom the camera that comes with the system and the images are pre-processed at first then normalization is done for the model to understand it better and the images are fed into the machine learning model to get the required output of hand landmarks. These landmarks represent which part of the hand it is and these landmarks can be used to calculate the distance between the fingers and can identify the hand gestures to perform the action. The speech recognition system already has some pre-defined actions and functions when triggered it the right keyword it can be used to perform a certain action, since both of them are customizable the user can keep personal actions as well as general actions.

2. LITERATURE REVIEW

M. Al-Hammadi [1] proposed a solution to develop a 3D CNN model that identifies hand regions and the input is at first processed, In addition to it, to reduce the unwanted features in the frame, cropping and spatial normalization is used to minimize the influence of it, and the hand region is cropped out and normalization is done to the image to focus more on the fingers configurations to detect the gestures. It makes use of the configurations that are local and global to effectively focus more on the hand region and get required features from it. There is no distraction since the unwanted features are removed using cropping techniques

G. Muhammad [2] proposed an method to develop a 3DCNN hand recognition model that focus more on the region rather than unwanted features that affect the model using various techniques like spatiotemporal features for hand gestures recognition and using a separate algorithm to normalize the structural dimensions of the gestures that comes from the video based on the facial position of the user. Spatial Dimension is also needed to control the constant change in distance and height from the camera. A Deep 3D CNN model is used to obtain the local spatio-temporal features that comes from the gestures sequence. The obtained features are given as input into the SoftMax layer for classification and fine-tuned by backpropagation.

W. Zhang[3] proposed an idea to combine a network that combines various different modules together to learn both short and long-term features from video inputs while preventing highprocessing or computation. A frame is selected randomly from each group and displayed as an RGB image and as a sequence of images. These two items are combined and input into a convolutional neural network to extract the features. The parameters for all groupings are shared by the CNN.

F. S. Khan [4] proposed an idea to develop a 3D Hand gestures recognition through Mask RCNN and classification of six different types of 3D hand recognition models with different parameters. It classifies hand motion with a per-trained ResNet 101 Deep Neural

Network model to classify hand gestures and Feature Pyramid Network (FPN) for extracting features.The Region Proposal Network (RPN) is used as substitute to a selective search process, since the production of bounding boxes using a selective procedure is extremely slow.

Z. Lu[5] proposed an idea to develop an effective tool for the extraction of discriminant features to perform gesture detection in Mobile Systems.A rational decision with distance measure is used to implement the one-shot learning of hand gesture recognition.To improve the proposed method's learning and classification performance, a mechanism of discrimination evolution with the innovation of new sample and voting integration based on multi-classifiers is established.

Similar to these,Many studies have been carried to solve human-computer interaction problems using machine learning, computer vision, and speech recognition systems. Like in [6], [8], [9] the Hand Recognition part is implemented using various techniques of computer vision and Kinects. The Study of hand-held systems [8] embraces that hand recognition is effective when each process is separated and has its own unique control. In [7] and [10], The Voice recognition part is studied and analyzed for better understanding and improvement of the Voice Command system. In [11], [12], [13], [14], [15] the study about audio enhancements for better speech recognition and computer vision techniques to use in processing to get at-most data from the hand gestures images have been studied

3. METHODOLOGY

This Project mainly focuses on giving people an alternative and a better method of interacting with the computer system. It uses novel technologies that are simple to use and not very complex, It consists of three major stages the Hand recognition part, the Voice recognition and command part, and then finally the computer interaction app which combines both the features and executes the command from a single place. Hand recognition and Voice commands part works independently and both of them are controlled by a single-core system which is triggered when a user runs or clicks that application.The applications have various step process that is executed in a step-by-step manner in order to run the application effectively. The Input Pre-Processing step is to get the hand images from the camera input and make them ready, In order to give it to the

Mediapipe Hand model.

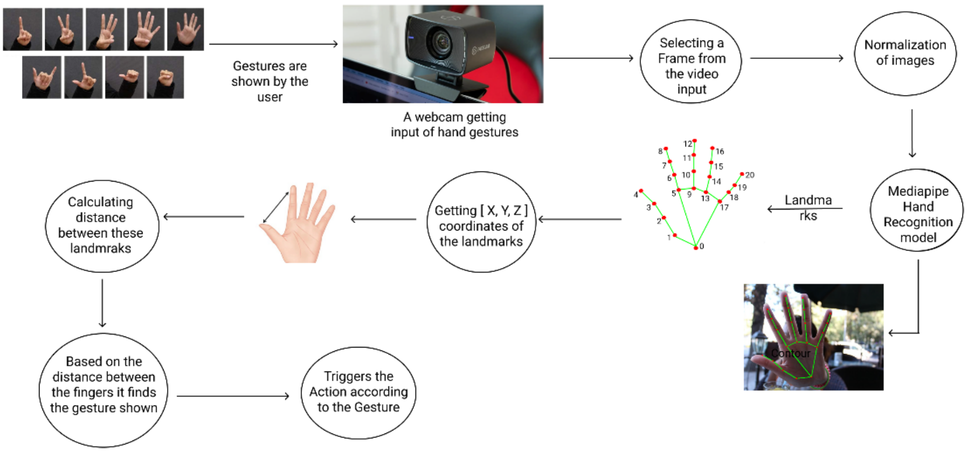

Gestures are shown by the user
A webcam getting input of hand gestures
Selecting a Frame from the video input
Normalization of images
Mediapipe Hand Recognition model
Landmarks
Getting [X, Y, Z] coordinates of the landmarks
Calculating distance between these landmraks
Based on the distance between the fingers it finds the gesture shown
Triggers the Action according to the Gesture
Contour

Figure3.1. Architecture of Hand Gesture Recognition

These Input videos are converted into a series of RGB frames, which are then converted to images taken from the video input.These images are normalized at first. This normalization is re-scaling the images to get better results. The re-scaling allows fairness of the input images and provides a better detection accuracy when using the stream of the images in the form of video. Atthe start of the application, the hand gestures recognition system is called and the application gets video input using OpenCV. The Video input is pre-processed at first and then it is fed into the Mediapipe hand recognition model.The Mediapipe hand recognition model is a high-resolution tracking solution for hands and fingers.It uses various Machine Learning algorithms to detect 21 different 3D Hand Markings from a single shot. It delivers real-time performance on a cell phone or desktop and it even scales up to multiple hands. Whereas the other machine learning solutions mostly rely on the hardware requirements.There are a total of 21 different landmarks for the hand starting from the wrist. These landmarks' Real-time locations are obtained from the media pipe hand recognition model. The obtained locations are an Array of size 21, each one of them containing x, y and z coordinates of that specific landmark.After obtaining these landmark locations, the labelingas done for each of the elements with their respective marking in the hand.Using these markings, It is easily accessible and more relevant. The distance calculation is done between all of these landmarks and the system checks for gestures shown and triggers the action for it. For a specific gesture, a set of landmarks is picked and a distance space is calculated based on the values the system understands what type of gesture it is and triggers the relevant action for it.

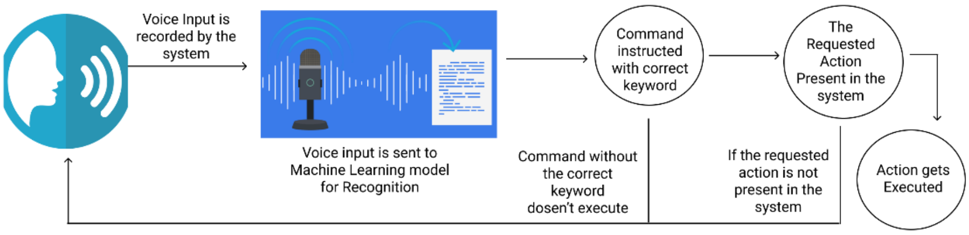
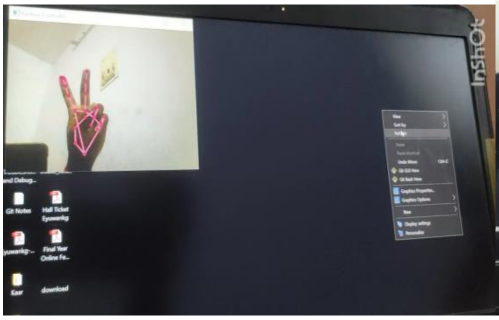

Voice Input is recorded by the system
Voice input is sent to Machine Learning model for Recognition
Command instructed with correct keyword
Command without the correct keyword doesn't execute
The Requested Action Present in the system
If the requested action is not present in the system
Action gets Executed

Figure 3.2. Architecture of Voice Command System

The Voice Commands gets the audio input from the system and process that data and trigger the action relevant to it. The voice commands are distinguished using a specific keyword that separates a normal speech and a command for the system. The keyword can be set by the user according to their convenience. The System check for the keyword when the user triggers for voice command system. When said with the right keyword, it takes the action that needs to be done and understands the context with it, and checks whether an action is specified for it, If there is an action specified for it, it just executes the function whichtriggers that action. Since all of this is voice input data, there can many actions which can be personalized for the different user according to their usability

4. RESULTS AND DISCUSSION

The User initializes the HCI-System in their computer and the gestures recognition, voice command module get permission for the camera and mic to capture the frame and voice of the user. The System receives the user's video feed of hand showing gestures and performs the action based on the gesture, this process is continuous as it happens in real-time. The Voice Command System gets the command with a keyword and checks whether the requested command action is defined by the user. If the Action is present in the system then the action gest executed



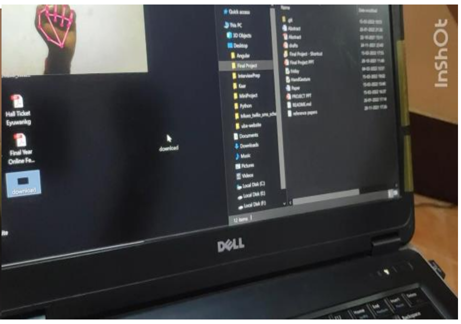
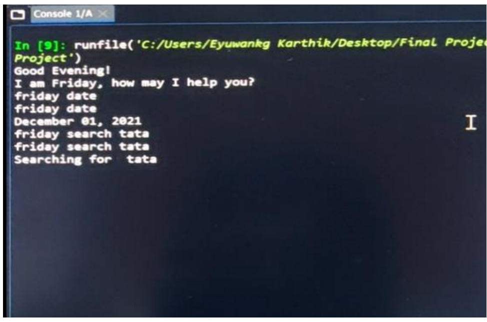
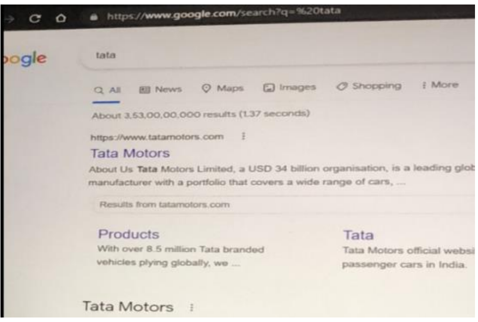


Figure4.1. Gesture for Right, Left Clicks System

Figure 4.2. Gesture for Drag-drop



```
In [9]: runfile('C:/Users/Eyuwankg Karthik/Desktop/Final Proje
Project')
Good Evening!
I am Friday, how may I help you?
friday date
friday date
December 01, 2021
friday search tata
friday search tata
Searching for  tata
```



https://www.google.com/search?q=%20tata

tata

About 3,53,00,00,000 results (1.37 seconds)

https://www.tatamotors.com

Tata Motors

About Us Tata Motors Limited, a USD 34 billion organisation, is a leading glob manufacturer with a portfolio that covers a wide range of cars, ...

Results from tatamotors.com

Products

With over 8.5 million Tata branded vehicles plying globally, we ...

Tata

Tata Motors official websi passenger cars in India.

Tata Motors

Figure 4.3. Voice Search Command Figure 4.4. Result for Voice Search Command

5. CONCLUSION& FUTURE SCOPE

The main aim of the project is to give an alternative and a better approach to communicate with a computer by adopting into a natural interaction system. There are various ways humans benefit from computers, but the traditional way of interacting with the computer is to have physical hardware of a mouse and keyboard, by eliminating that we have a device that is very much portable and has more space for other improvements in hardware components. Especially some disabled people cannot interact with the computer the usual way like people who lost a hand, for them, it is very useful for interaction and controlling the system. The Future implementation of this project is to add more gestures and have the option of two hands for the user to give as gestures and add more actions to voice commands.

6. REFERENCES

[1] M. Al-Hammadi et al., "Deep Learning-Based Approach for Sign Language Gesture Recognition With Efficient Hand Gesture Representation," in IEEE Access, vol. 8, pp. 192527-192542, 2020, doi: 10.1109/ACCESS.2020.3032140.

[2] G. Muhammad ,M. Al-Hammadi, W. Abdul, M. Alsulaiman, M. A. Bencherif and M. A. Mekhtiche, "Hand Gesture Recognition for Sign Language Using 3DCNN," in IEEE Access, vol. 8, pp. 79491-79509, 2020, doi: 10.1109/ACCESS.2020.2990434.

[3] W. Zhang, J. Wang and F. Lan, "Dynamic hand gesture recognition based on short-term sampling neural networks," in IEEE/CAA Journal of AutomaticaSinica, vol. 8, no. 1, pp. 110-120, January 2021, doi: 10.1109/JAS.2020.1003465.

[4] F. S. Khan, M. N. H. Mohd, D. M. Soomro, S. Bagchi and M. D. Khan, "3D Hand Gestures Segmentation and Optimized Classification Using Deep Learning," in IEEE Access, vol. 9, pp. 131614-131624, 2021, doi: 10.1109/ACCESS.2021.3114871.

[5] Z. Lu, S. Qin, L. Li, D. Zhang, K. Xu and Z. Hu, "One-Shot Learning Hand Gesture Recognition based on Lightweight 3D Convolutional Neural Networks for Portable Applications on Mobile Systems," in IEEE Access, vol. 7, pp. 131732-131748, 2019, doi: 10.1109/ACCESS.2019.2940997.

[6] N. Mohamed, M. B. Mustafa and N. Jomhari, "A Review of the Hand Gesture Recognition System: Current Progress and Future Directions," in IEEE Access, vol. 9, pp. 157422-157436, 2021, doi: 10.1109/ACCESS.2021.3129650.

[7] R.Bharathi, T.Abirami,” Energy efficient compressive sensing with predictive model for IoT based medical data transmission”, Journal of Ambient Intelligence and Humanized Computing, November 2020, https://doi.org/10.1007/s12652-020-02670-z

.

[8] S. Kim, G. Park, S. Yim, S. Choi and S. Choi, "Gesture-recognizing hand-held interface with vibrotactile feedback for 3D interaction," in IEEE ransactions on Consumer Electronics, vol. 55, no. 3, pp. 1169-1177, August 2009, doi: 10.1109/TCE.2009.5277972.

[9] C. Jeon, O. -J. Kwon, D. Shin and D. Shin, "Hand-Mouse Interface Using Virtual Monitor Concept for Natural Interaction," in IEEE Access, vol. 5, pp. 25181-25188, 2017, doi: 10.1109/ACCESS.2017.2768405.

[10] H. Lee, S. Chang, D. Yook and Y. Kim, "A voice trigger system using keyword and speaker recognition for mobile devices," in IEEE Transactions on Consumer Electronics, vol. 55, no. 4, pp. 2377-2384, November 2009, doi: 10.1109/TCE.2009.5373813.

[11] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 3, pp. 1108-1124, 1 May-June 2021, doi: 10.1109/TDSC.2019.2906165.

[12] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 3, pp. 1108-1124, 1 May-June 2021, doi: 10.1109/TDSC.2019.2906165.

[13] C. Shi, D. Yang, J. Zhao and H. Liu, "Computer Vision-Based Grasp Pattern Recognition With Application to Myoelectric Control of Dexterous Hand Prosthesis," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 9, pp. 2090-2099, Sept. 2020, doi: 10.1109/TNSRE.2020.3007625.

[14] D. Połap and M. Woźniak, "Voice recognition by neuro-heuristic method," in Tsinghua Science and Technology, vol. 24, no. 1, pp. 9-17, Feb. 2019, doi: 10.26599/TST.2018.9010066.

[15] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif and M. A. Mekhtiche, "Hand Gesture Recognition for Sign Language Using 3DCNN," in IEEE Access, vol. 8, pp. 79491-79509, 2020, doi: 10.1109/ACCESS.2020.2990434.