ISSN No: 2456-2156

Control Desktop Applications with a Simple Webcam and Gesture Recognition using DL

Smit Parikh
Computer Science and Engineering
(Network and security) MIT-ADT University
Pune, India

Srikar Banka
Computer Science and Engineering
(Network and security) MIT-ADT University
Pune, India

Isha Lautrey
Computer Science and Engineering
(Network and security) MIT-ADT University
Pune, India

Isha Gupta
Computer Science and Engineering
(Network and security) MIT-ADT University
Pune, India

Abstract:- In recent years, hand gesture recognition has been used in a variety of fields, especially in the area of man-machine interaction (MMI), where it is regarded as a more natural and versatile input than conventional input devices such as mice and keyboards. Since there is a high distance between the user and the machine, using a physical controlling device such as a keyboard and mouse for human interaction with the computer hinders the normal interface. Our goal is to solve this problem by developing an application that uses hand movements to monitor some of the basic computer functions through an integrated webcam. To make our tasks easy, a Hand Gesture Recognition device senses gestures and converts them to specific actions. With the aid of the Jester dataset, a model can be created using a 3D convolutional neural network and deep learning, which will be interfaced using Django, React.JS, and Electron. The key outcome predicted is that the user, using hand gestures, would be able to monitor the system's basic functions, providing them with the ultimate convenience.

Keyword:- *Human-Computer Interaction, Jester Dataset, 3D Convolutional Neural Network, Deep Learning.*

I. INTRODUCTION

Deep neural network:-

A deep neural network (DNN), or deep net for short, is a neural network with a certain degree of complexity, typically at least two layers. Deep nets use advanced math modeling to process data in complex ways.

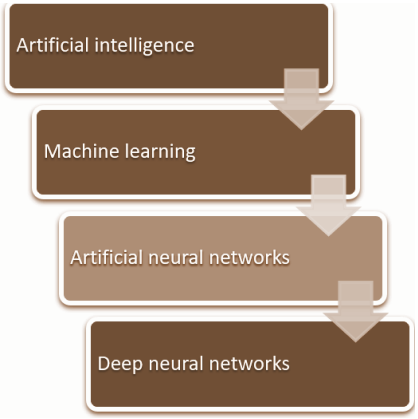

Figure 1 - Deep Learning

Before deep nets, a few things had to be built (Fig. 1): Machine learning had to be established first. ML is a system for automating (via algorithms) statistical models, such as linear regression models, to improve prediction accuracy. A single model that makes predictions about something is referred to as a model. Those forecasts are reasonably accurate. A learning algorithm (machine learning) takes all of its incorrect predictions and adjusts the weights inside the model to construct a model that produces fewer errors. Artificial neural networks arose from the learning portion of the modeling process. The hidden layer is used by ANNs to store and determine how important each of the inputs is to the output. The secret layer stores knowledge about the value of an input and creates correlations about the importance of input combinations.

ISSN No: 2456-2156

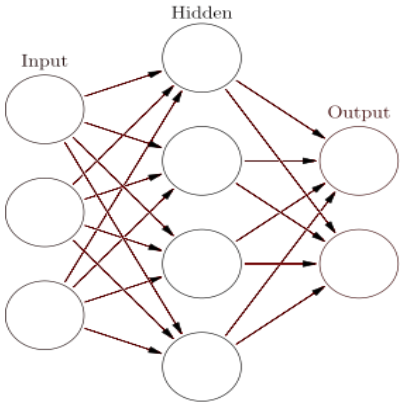

Figure 2 - DNN Working

A simple working of DNN is represented in Fig. 2. Deep neural nets, on the other hand, make use of the ANN portion. They claim that if that works well at improving a model—because a node in the hidden layer makes all connections and grades the value of the input to decide the output—then more and more of these are piled on top of each other, maximizing the hidden layer's benefits.

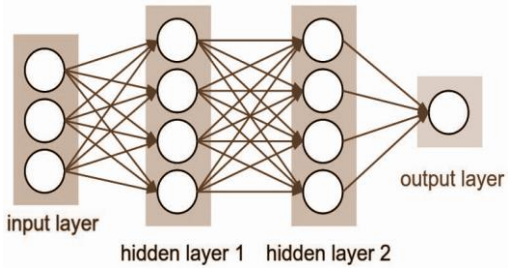

Figure 3 - Hidden layers in DNN

The term 'deep' refers to a model's layers having several layers. As a result, there are several hidden layers in the deep net. A multiple hidden layer structure is shown in Fig. 3.

Convolutional Neural Networks:-

A convolutional neural network (CNN) is a form of artificial neural network that is specifically designed to process pixel data and is used in image recognition and processing. It's a feed-forward neural network of up to 20 or 30 layers in most cases. The convolutional layer is a special type of layer that gives a convolutional neural network its strength.

CNNs are image processing, artificial intelligence (AI) systems that use deep learning to perform both generative and descriptive tasks. They frequently use computer vision, which includes image and video recognition, as well as recommender systems and natural language processing (NLP). Many convolutional layers are placed on top of each other in convolutional neural networks, each capable of recognizing more complex shapes. Handwritten digits can be recognized with three or four convolutional layers, and human faces can be distinguished with 25 layers.

A CNN employs a device similar to a multilayer perceptron that is optimized for low processing requirements. An input layer, an output layer, and a hidden layer with several convolutional layers, pooling layers, completely linked layers, and normalization layers make up CNN's layers.

The elimination of limitations and improvements in image processing performance result in a device that is both more efficient and easier to train for image processing and natural language processing.

In a convolutional neural network, the use of convolutional layers mimics the structure of the human visual cortex, where a sequence of layers process an incoming image and recognize increasingly more complex features.

Jester dataset:-

The 20BN-JESTER dataset contains a large number of classified video clips of people making pre-determined hand gestures in front of a laptop camera or webcam. A large number of crowd workers contributed to the dataset's development. It enables the creation of robust machine learning models capable of recognizing human hand gestures.

Human computer interaction:-

Human-computer interaction (HCI) is a multidisciplinary area of research that focuses on computer technology design and, in particular, the interaction between humans (users) and computers. HCI began with computers and has since grown to include almost all aspects of information technology design.

As computers became less costly, room-sized resources designed for experts in specialized environments, the need for human-computer interaction that was also simple and efficient for less experienced users grew. HCI has grown to include a variety of disciplines, including computer science, cognitive science, and human-factors engineering, since its founding.

HCI was soon the subject of a lot of academic research. Those who trained and worked in human-computer interaction (HCI) saw it as a critical tool for popularising the concept that computer-user interaction could mimic a human-to-human, open-ended conversation. HCI researchers initially concentrated on improving the usability of desktop computers (i.e., practitioners concentrated on how easy computers are to learn and use). However, as new technology such as the Internet and smartphones become more widely available, computer use will shift away from the desktop and toward handheld devices.

Gesture recognition:-

The hand gesture recognition device is used to interact between a computer and a human using hand gestures. A windows-based program is developed for live motion gesture recognition using a webcam.

ISSN No: 2456-2156

This project is a combination of live motion detection and gesture recognition.

This software makes use of the webcam to detect a user's motion and perform basic operations in response. A simple gesture must be made by the person. After that, the webcam tracks the user's motion and detects the gesture. It then recognizes the user's gesture (by comparing it to a collection of predefined gestures) and performs the appropriate action. This application can be made to run in the background while the user is using other programs and applications. For a hands-free method, this is extremely useful.

Though it may not be very useful for searching the web or writing a text document, it is extremely useful in media players and also when reading documents or files. Even if you're not in front of the monitor, a simple gesture will pause, play, or increase the volume. Also, one could easily scroll through an eBook or a presentation while eating lunch.

There are a variety of applications for gesture recognition, including:

- Socially assistive robotics
- Virtual control
- Remote control
- Control through facial gesture
- Aid to the physically challenged
- Immersive game technology
- Sign language recognition

II. LITERATURE REVIEW

In 2020, Munir Oudah [1] has stated that hand gestures can be used in several fields such as clinical and health, Sign Language Recognition, Robot Control, Virtual Environment, Personal Computer and Tablet, Home Automation and Gaming. Hand signals were categorised under a variety of headings, including stance and motion, as well as dynamic and static, or a combination of the two. The paper reviews the literature on hand gesture techniques and states pros and cons under different circumstances. It evaluates the performances through techniques of Computer Vision that deal with the similarity and difference points, technique of hand segmentation along with classification algorithms. It's experimented using multiple hand gesture methods like Hand Gestures Based on Instrumented Glove Approach, Hand Gestures Based on Computer Vision Approach (Color-Based Recognition, Appearance-Based Recognition, Motion-Based Recognition etc.). The paper covers all aspects like drawbacks, number and types of gestures, dataset used, detection range (distance) and type of camera used and A broad overview of hand gesture approaches is presented, along with a review of some potential applications.

In 2020, Vaidyanath Areyur Shanthakumar [2] has proposed a novel angular velocity method that uses a sensor-based motion tracking system to capture 3D hand and finger motions and also to detect and recognize hand gestures. As it is directly applied to real-time 3D motion data, streamed by the sensor-based system, the approach is capable of recognizing both static and dynamic gestures in real-time. In this paper, the recognition accuracy and execution performance are assessed with interactive applications like “Happy Ball Game” that require gesture input to interact with the virtual environment. The results show high recognition accuracy(97.3%) , high execution efficiency (60 frames per second), and high-levels of usability and acceptance(above 90%).

In 2020, Dinh-Son Tran [3] has proposed a novel method for fingertip detection and hand gesture recognition in real-time using an RGB-D camera and a 3D convolution neural network (3DCNN). The system is capable of accurately and robustly extractracting fingertip locations and recognizing gestures in real-time. The accurateness and robustness of the interface by evaluating hand gesture recognition across a variety of gestures is demonstrated. The paper focuses on Hand Region Extraction, Fingertip Detection, Hand Gesture Spotting and Hand Gesture Recognition (3DCNN). It showed a high speed of tracking process with 30 frames per second. It is considered to be a good approach as the results not only showed a high level of accuracy of hand gesture recognition but is also suitable for practical applications. The proposed method has many advantages, for example, working well in changing light levels or with complex backgrounds, accurate detection of hand gestures at a longer distance, and recognizing the hand as it is a promising technique which can work well in real-time.

In 2020, Smit Parikh [4] tackled the problem of making an application that controls some specific functionalities of computers using hand gestures via an integrated webcam. The paper proposed a hand gesture recognition system for detecting gestures and translating them into specific actions for the sole purpose of making work easier. The paper states the pursuing of this system through OpenCV to capture the gestures which will be interfaced using Django, React.Js and Electron. The algorithm used to train the system is YOLO and the gesture saving will take place in the database (MongoDB). The paper has portrayed the survey related to hand gesture recognition, using it in different application areas and also studying existing solutions and majorly aims to provide the system’s user utmost ease and comfort.

In 2019, Joanna Materzynska [5] has showcased the largest collection of short clips of videos of humans performing hand gestures. The dataset is a collection gathered using the help of over 1300 different actors in their unconstrained environments. The baseline network proposes a 3D convolutional neural network (3D-CNN) that uses spatio-temporal filters as the main building block. The model achieves 93.87% of the top 1 accuracy. The submission platform allows users to test and compare their models across the latest state-of-the-art video recognition systems. The paper suggests a 3D CNN baseline model and shows that the vast amount of data offered to the computer vision community significantly impacts the performance of

ISSN No: 2456-2156

the network, which may explain the high accuracy scores on the leaderboard.

In 2021, Wenjin Zhang [6] has focused on a novel deep learning network for hand gesture recognition that integrates several well-proven modules together to learn both short-term and long-term features from video inputs. Two datasets are used here namely Jester and Nvidia. A frame is randomly selected and represented as an RGB image as well as an optical flow snapshot and these are fused and fed into a convolutional neural network (ConvNet) for feature extraction. The loss of the model has been on a downward trend, which indicates that the training process is effective. The paper also states that the ConvNets can successfully track the hand movements like "pulling two finger", "swiping right", and "stop sign".

In 2020, Bin Yu [7] has proposed an attentive feature fusion framework for efficient hand-gesture recognition. The model in this paper utilizes a shallow two-stream CNNs to capture the low-level features from the original video frame and its corresponding optical flow, The training and testing is done on a large-scale dataset called Jester and the performance results in 95.77% of classification accuracy. . For the future work, consideration is towards a late feature fusion strategy for fusing the information from RGB and optical flow at high-level instead of in an early low-level input feature fusion. To investigate the properties of the proposed method for hand gesture recognition, two principal characters are evaluated - the influence of number of segments and the effectiveness of using optical flow.

In 2019, Yang Yi [8] has focused on the 1D convolutional neural networks and proposed a simple and efficient architectural unit, Multi-Kernel Temporal Block (MKTB), that models the multi-scale temporal responses by explicitly applying different temporal kernels. Then, they present a Global Refinement Block (GRB), which is an attention module for shaping the global temporal features based on the cross-channel similarity.Using these two the architecture can effectively explore the spatiotemporal features within tolerable computational cost. The proposed MKTB and GRB are plug-and-play modules and the experiments on other tasks, like video understanding and video-based person reidentification, also display their good performance in efficiency and capability of generalization.

Year [Citation]	Methodology	Features	Challenges
2020 [1]	Recognition of Hand Gestures Based on techniques such as computer vision, hand segmentation, and classification algorithms.	In contrast to the sensor data glove, one of the methods, Color-Based Recognition using Glove Marker, is simple to use and inexpensive.	Problems with techniques in the interaction system like the Open-NI library or OpenCV, as well as the Dataset Glove approach, include illumination variation, accuracy difference, and so on.
2020 [2]	A sensor-based motion tracking device that uses an angular velocity approach to capture 3D hand and finger movements.	Both static and dynamic movements can be recognized in real-time using this method. This experimental results show accuracy of 93.7%	Challenges are minimal display space, deep and complex sorting menus, and a lack of technical support to take inputs
2020 [3]	Using an RGB-D camera and a 3DCNN, a fingertip detection and hand gesture recognition system is used in real-time (3DCNN).	The system shows a high level of accuracy up to 92.6% along with working well in different Environment and multiple hand gesture	Concerning 3D dense mapping, RGB-D sensors have major disadvantages, including limited measurement ranges (e.g., within 3 m) and errors in in-depth measurement increasing with distance from the sensor.
2020 [4]	Using OpenCV and YOLO to train a Hand Gesture Recognition system.	Great speed as it can process around 45 frames per second and understands generalized object representation.	In comparison to CNN, the system has a lower recall and higher localization error, fails to detect near objects because each grid can only suggest two bounding boxes, and also struggles to detect small objects.
2019 [5]	A simple 3D convolutional neural network is used to recognise gestures in the Jester Dataset.	The majority of the submissions were able to achieve 90% accuracy, which was aided by their dataset's large amount of training data.	The challenge is to gather maximum submissions, only 59 have been gathered out of which 3 are from the team itself.
2021 [6]	Short-term sampling neural networks are used in dynamic	An augmented dataset with increased diversity of hand gestures	Applying the model to untrimmed videos by sampling at a fair rate is expensive

	hand gesture recognition to learn both short-term and long-term features. Two datasets, Jester and Nvidia, are used to evaluate the model.	was used to demonstrate the model's robustness. On the Jester dataset, the average accuracy was 95.73 percent, and on the Nvidia dataset, it was 85.13 percent.	with the latest optical flow algorithm.
2020 [7]	Using the Jester dataset, hand gesture recognition is based on attentive feature fusion, which uses shallow two-stream CNNs to capture low-level features from the original video frame and its subsequent optical flow.	This model achieves better results at a lower computational cost. The method's efficiency improvements are due to the motion information captured by computing optical flow and attention-based feature selection, which has a 95.77 percent accuracy.	Instead of early low-level input feature fusion, a late feature fusion strategy for fusing the information from RGB and optical flow at a high level is being considered as a challenge.
2019 [8]	High-performance gesture recognition using the 1D convolutional neural networks by suggesting Multi-Kernel Temporal Block (MKTB) and Global Refinement Block (GRB).	MKTB uses several 1D depthwise convolutions to capture both short and long-term temporal information. MKTB and GRB retain the same size between input and output and can be easily deployed anywhere.	More research into the efficacy of the MKTB and GRB modules on more video comprehension tasks is the challenge to be faced.

Table 1:- Literature Review

III. PROPOSED SYSTEM

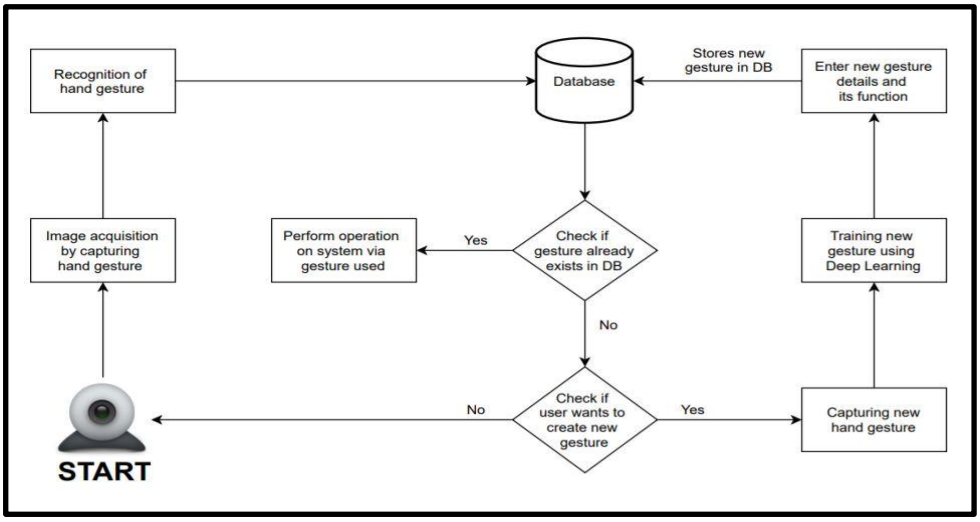

Figure 4 - Block Diagram

- Users of the existing systems have difficulty managing their laptops or desktops when they are in the midst of an activity.
- As a result, we've proposed a method for controlling specific system functionalities with hand gestures. The workflow of our proposed system is depicted in the block diagram above. Our proposed approach, like most other gesture recognition systems, has three main phases.
- For security purposes, the first stage of this proposed approach illustrates the development of a web application with two-factor authentication. Thereafter, the user's hand gesture is captured and recorded. A few predefined hand movements will be provided so that the user can get a sense of how to control the computer.
- Differentiated features and effective classifier selection are major issues in most of the studies. To avoid this issue, 3D-CNN is used to build a model, and a Deep Learning algorithm is used to train the model using the Jester dataset.
- The second stage determines what kind of hand gesture was captured. The hand gesture is then checked to see if it already exists in the database. If it doesn't exist, the user is given the option of returning to the first phase of image acquisition from the camera or creating a new hand gesture for that specific function. The user's new hand gesture is saved in the database after it gets created.
- Following the addition of the latest hand gesture, the third stage is initiated. Now, the user may perform the operation that he or she has created, thereby allowing

him/her to control the computer without having to touch it physically. (For instance, play, pause, mute, or open a program).
- Not only static but also dynamic hand gestures can be saved.

IV. RESULTS

In the checkpoint figure (Fig. 5) represents checkpoints where our model data, that has been trained by using jester dataset, is stored in the main model file.

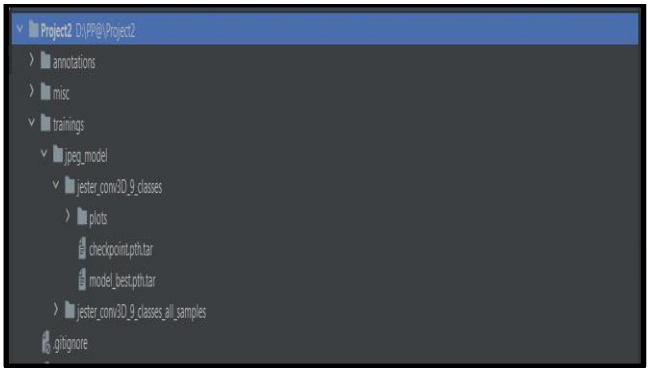

Figure 5 - Checkpoint

During the training process of the model, a training accuracy of 99.36% is observed while during the testing process of the model, a testing accuracy of 95.33% is observed in the accuracy training and accuracy testing figure (Fig. 6).

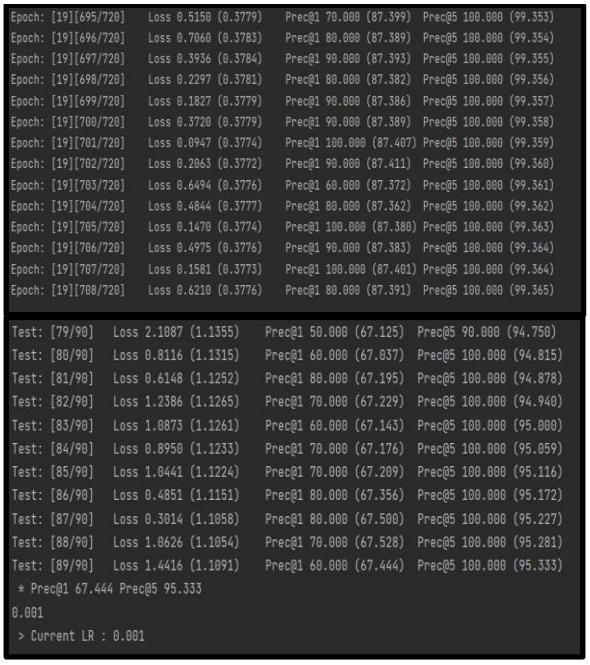

Figure 6 - Train and Test Accuracy

In the doing other things figure (Fig. 7), when the webcam detects that the user is not using any hand gestures and he/she is dynamic at that very moment, then a 'Doing other things' message is shown.

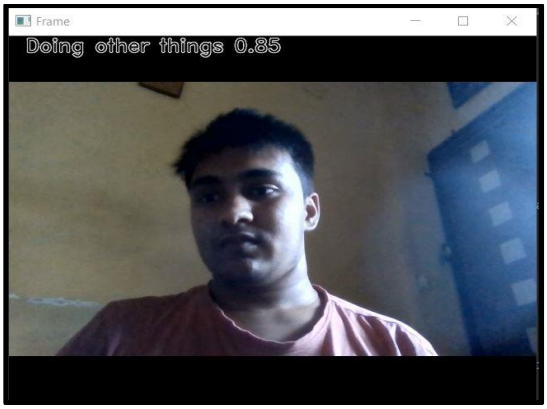

Figure 7 - Doing other things

The no gesture figure (Fig. 8) depicts when the webcam detects that the user is not using any hand gestures and he/she is static at that very moment, then a 'No gesture' message is displayed.

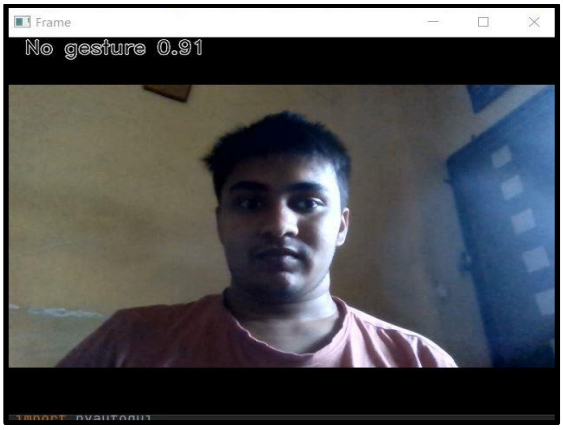

Figure 8 - No Gestures

The stop sign figure (Fig. 9) displays a 'Stop sign' message when it detects a whole palm of the user, the video is paused/played accordingly in the vlc media player.

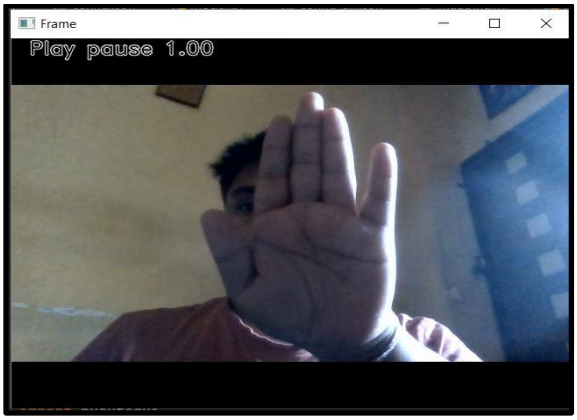

Figure 9 - Stop Sign

In the swipe right figure (Fig. 10), when the user simply swipes right with his index finger which indicates forwarding of the video by 10 seconds in the vlc media player, a 'Swipe Right' message is displayed.

ISSN No: 2456-2156

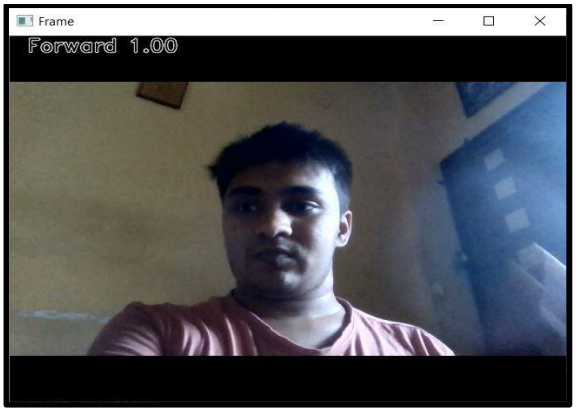

Figure 10 - Swipe Right

The swipe up figure (Fig. 11) shows a 'Swipe up' message when the user swipes his/her index finger upwards which causes an increase in the volume of the video by 5 units.

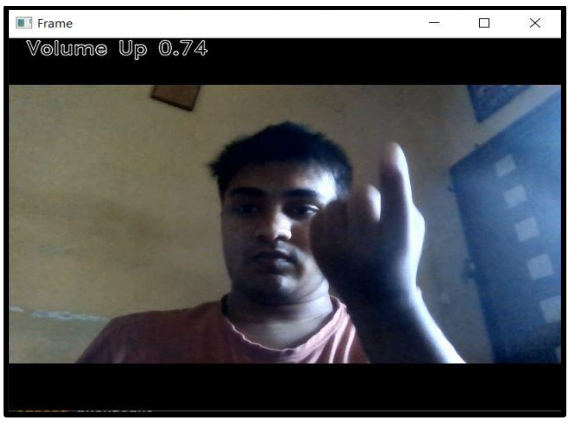

Figure 11 - Swipe Up

V. CONCLUSION

Every day, technology advances and progresses to make a man's life and work much simpler and more comfortable. We've taken this into account, and we've conducted research into using hand gesture recognition to control specific computer system features.

This paper has presented a survey on hand gesture recognition, its implementation in various fields, and a review of current solutions. Based on our research, we've proposed an innovative way to control our systems by recognizing hand gestures, which will provide the user with more convenience and comfort.

REFERENCES

[1]. Hand Gesture Recognition Based on Computer Vision: A Review of Techniques by Munir Oudah, Ali Al-Naji, and Javaan Chahl in 2020.
[2]. Design and evaluation of a hand gesture recognition approach for real-time interactions by Vaidyanath Areyur Shanthakumar, Jeff Hansberger, Lizhou Cao, Sarah Meacham, Victoria R. Blakely and Chao Peng in 2020.
[3]. Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network by Dinh-Son Tran, Ngoc-Huynh Ho, Hyung-Jeong Yang, EuTteum Baek, Soo-Hyung Kim and Gueesang Lee in 2020.
[4]. Human-Computer Interaction using Dynamic Hand Gesture Recognition to conveniently control the system by Smit Parikh, Srikar Banka, Isha Lautrey and Isha Gupta in 2020.
[5]. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures by Joanna Materzynska, Guillaume Berger, Ingo Bax and Roland Memisevic in 2019.
[6]. Dynamic Hand Gesture Recognition Based on Short-Term Sampling Neural Networks by Wenjin Zhang, Jiacun Wang, Senior Member and Fangping Lan in 2021.
[7]. Hand gesture recognition based on attentive feature fusion by Bin Yu, Zhiming Luo, Huangbin Wu and Shaozi Li in 2020.
[8]. High Performance Gesture Recognition via Effective and Efficient Temporal Modeling by Yang Yi, Feng Ni, Yuexin Ma, Xinge Zhu, Yuankai Qi, Riming Qiu, Shijie Zhao, Feng Li and Yongtao Wang in 2019.
[9]. Enhancing User Experience Using Hand - Gesture Control by P. Sai Prasanth, Aswathy Gopalakrishnan, Oviya Sivakumar and A. Aruna in 2019.
[10]. Design of control system based on hand gesture recognition by Shining Song, Dongsong Yan and Yongjun Xie in 2018.
[11]. Light YOLO for High-Speed Gesture Recognition by Zihan Ni, Jia Chen, Noang Sang, Changxin Gao and Leyuan Liu in 2018.
[12]. Motion feature network: Fixed motion filter for action recognition by M. Lee, S. Lee, S. Son, G. Park, and N. Kwak in 2018.
[13]. Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. by Y. Zhang, C. Cao, J. Cheng, and H. Lu in 2018.
[14]. Temporal relational reasoning in videos by B. Zhou, A. Andonian, A. Oliva, and A. Torralba in 2018.
[15]. Aggregated residual transformations for deep neural networks by S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He in 2017.
[16]. A new model and the kinetics dataset by J. Carreira and A. Zisserman in 2017.
[17]. Gesture Recognition and fingertip detection for Human Computer Interaction by R. Meena Prakash, T. Deepa, T. Gunasundari and N. Kasthuri in 2017.
[18]. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis by S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. in 2017
[19]. Smart gloves for hand gesture recognition: Sign language to speech conversion system by K. A. Bhaskaran, A. G. Nair, K. D. Ram, K. Ananthanarayanan, and H. N. Vardhan in 2016.
[20]. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network by P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz in 2016.

[21]. convolutional networks for action recognition in videos by K. Simonyan and A. Zisserman. Two-stream in 2014.
[22]. Chairgest: a challenge for multimodal mid-air gesture recognition for close hci. by S. Ruffieux, D. Lalanne, and E. Mugellini in 2013.
[23]. Imagenet classification with deep convolutional neural networks. by A. Krizhevsky, I. Sutskever, and G. E. Hinton in 2012.