INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, MARCH 2026 | ISSN: 2456-4184 | IJNRD.ORG

Deep Learning Based Media Control With Gesture Recognition

Ms.N.Soujanya
Department of CSE(Data Science)
CMR Technical Campus Hyderabad,India

1st TammaliMadan
Department of CSE(Data Science)
CMR Technical Campus Hyderabad,India

2nd Banoth Raj Kumar
Department of CSE(Data Science)
CMR Technical CampuHyderabad, India

3rd Ganjay Karthik
Department of CSE(Data Science)
CMR Technical CampusHyderabad, India

Abstract—This study proposes an automated Deep LearningBased Media Control System that can be used to substitute the traditional manual input methods in modern computing environments. The Deep Learning based media control with gesture recognition uses OpenCv and MediaPipe. Where OpenCv is used to control media easily without touching any device physically. MediaPipe is used to track movements of human hands. In this system users just need to show simple gestures and then the system recognizes exact actions like Pause, Play, Volume Up, Volume Down, Next, Previous. This system takes the detail hand gesture and decides which action to perform. It stores previously performed actions and guesses the gestures in less time. Overall, this system improves the accuracy and reduces physical touch. The Convolutional neural network is one type of deep learning model which is good at understanding and processing images. The main purpose of using CNN in deep learning based media control with gesture recognition is to automatically extract features from hand gestures, understand finger position and hand shapes, and work well even in less lighting and background. PyAutoGUI is one type of python library which acts as a bridge between recognized gestures and actual system gestures. Once a gesture is performed in front of the camera the PyAutoGUI makes the connection between recognized gesture and keyboard actions.

Keywords- MediaPipe, OpenCv, Deep Learning, Gestures, accuracy.

I. INTRODUCTION

Media controlling is important in daily life in computer systems. Media controlling helps users to adjust or control in an easy and comfortable manner. Controlling media through traditional methods, such as Remote Controls, Keyboards and Mouse Clicks can be difficult when the user is multitasking. It can be a problematic situation for users. To overcome these issues, there has been new development adopting the use of gesture recognition technology to enable automation of media control. Indicatively, Sharma highlighted the effectiveness and comfort of a smart gesture-based interaction framework that operated on hand- tracking models to minimize the physical effort handled by a user and eliminate the dependence on traditional input tools [1]. On the same note, contemporary systems are becoming inclined toward digital and AI enhanced interfaces to enhance operational accuracy and intuitiveness. Gesture-based control systems use deep learning and computer vision models to capture hand movements, map them to structured features, and identify the intended gesture in real time. According to Kumar et al., gesture recognition and interactive frameworks can indeed be coupled to boost the concept of accessibility and realtime responsiveness by adding pipeline command execution to merge reflexes [2]. Patel et al. also discovered that offering camera-based gesture control does not necessitate specific hardware and is also broad-range in noncontact interaction of the system utilisation [3]. With these new things, AIenabled gesture systems are on the path towards more scalable, less intrusive and faster human- computer interface.

The major concerns when evaluating a powerfully appropriate gesture-based automation framework in the computing scenario are precision, strength, and user-friendliness. Lari et al. highlighted that machine-learning-enhanced hand landmark detection improves dependability and reduces false gesture activations across different environments [4]. Meanwhile, Bangare et al. confirmed that automated gesture-based media systems increase operational efficiency and reduce dependency on physical input devices [5]. Decision support modules and optimized recognition models also add value by improving adaptability and flexibility in detecting varied user gestures under different viewing angles and lighting conditions [6]. Thus, a deep-learning- based gesture recognition system can help achieve significant benefits for users seeking secure, efficient, and user- friendly media control solutions .

II. LITERATURE SURVEY

Camera-Based Gesture Recognition Techniques

In earlier days the gesture recognition system was created using some hardware resources like data gloves, infrared devices and wearable sensors. These tools are used to measure movements of human hands accurately, but this leads to discomfort because users need to wear equipment. As users want an easier way to interact with the computer, the researchers innovatively started using cameras instead of hardware resources like wearable sensors etc. This project mainly focuses on using web cameras to detect gestures through landmarks on the user 's hand. It does not need any physical devices. Through this users may feel comfortable and friendly.

Advanced improvement came in the year of 2000 through "Bradski" introducing OpenCv, an open source.

Vision library. It is used to process real-time images faster and easier. Main functioning of OpenCv is, it provides some tools to use for detecting hand landmarks, hand shapes, finding edges and detecting motion. Deep learning mainly helped gesture recognition systems to work well in different lightings and backgrounds. In today's world, we can see Camera-Based

Gesture recognition is used everywhere some of them are in Smart homes, Gaming, Robotics and touch-free controls.

Landmark technique is one of the most advanced techniques where it is used to detect hand parts such as finger tips, joints and the wrist. One of the well known techniques is "Google's Media Pipe hand" which can detect 21 hand landmarks. It works in a step by step process like finding the palm, selecting the hand area, and finding the landmark position.

A. Landmark Detection Techniques for Hand Tracking

Landmark-based hand tracking is a modern technique mainly used to detect hand movements accurately. This technique mainly works by identifying important points on the hand, such as fingers, fingertips, finger points, and the wrist. The most popular system used in this project is MediaPipe hands. It is designed to identify 21 major landmarks of a hand with further precision. The method is more effective as both mobile gadgets and heavy computers can employ it. It is also effective in various conditions and backgrounds, which aid in the working in a real environment. The technique is enjoying numerous benefits since it can be applied in numerous ways including sign language, robotics, augmented and virtual. contactless used interfaces, reality, and reality.

B. Deep Learning Approaches for Gesture Classification

A seminal contribution by Krizhevsky et al. (2012) in the form of Convolutional Neural Networks (CNNs) suggested a new potent model that was utilised in a large scale image classification and thus resulted in the adoption of deep learning in gesture recognition tasks on a mass scale. The reason behind using Convolutional Neural Network is that it is capable of finding significant elements of images and videos without human intervention.

Its primary application is in correct determination of gestures. It is also possible to find some popular deep learning models like ResNet, as well as MobileNet that is primarily configured in the control of Gesture- based models. These models are desirable in the process of deriving features of images and videos. The works of Bengio, and Courville contribute to the knowledge of deep learning and demonstrate us why it is effective in solving the complicated tasks. Due to such accuracy and flexibility, CNN-based systems are highly adopted in real-time interaction systems. They are also convenient method of scaling, adjusting to new surroundings and working in the actual conditions. These methods are mostly applied in Gesture-Based Systems to follow the movements of hands and fingers shape.

C. Hybrid Real-Time Gesture Recognition Systems

The recent work in Gesture-based recognition mainly focuses on systems that are having the mix of traditional computer vision methods with more advanced deep learning models to increase accuracy, speed and stability. These systems combine technologies like MediPipe, OpenCV, and CNN(Convolutional Neural Network) where each method has its own way of supporting the process of Gesture Recognition. The MediPipe mainly focuses on identifying 21 key landmarks on human hands, such as finger tips, finger shape, finger joints. This works well in low lighting, low background conditions. The recent research studies, including work by saha and basu, shows that the combined system works well in applications like robotics, smart systems, and virtual spaces.

D. Gesture-Based Human–Computer Interaction Applications

The recent researcher Margaryan, and many researchers studied that Gesture-based control system is a useful way in which it helps to improve how users interact with digital devices. The Gesture-Based Human-Computer Interaction (HCI) is becoming more popular because it gives access to people to interact with computers in a moderate and easy way, by reducing the need of keyboards or mouse. Gesture interactions mainly used in applications like gaming, virtual and robotics.

III. PROPOSED METHODOLOGY

A. General Prospectus of the Projected System

The proposed system enables completely contactless media control by interpreting hand gestures through deep learning and computer vision. The main objective of this system is to allow users to interact with the media players without using physical input devices such as keyboards, mice, or remote controls. Instead, a normal webcam is used to capture hand movements, which is used to analyze and transform into media control commands.

After extracting the hand features, gesture recognition is performed. Simple gestures are identified using predefined rule based logic, more complex gestures are recognized by using a Convolutional Neural Network(CNN) model. This combination improves accuracy and allows the system to handle different types of gestures effectively. After the gesture has been registered it is correspondingly mapped to a given set of medial control operations, like play, pause, previous track, next track, and volume increase/volume decrease. PyAutoGui is used in executing these actions. It allows connecting to the media player directly. This whole system is offline and does not need any special hardware and this makes it low cost and simple to operate. This system is particularly applicable to the old population as well as the disabled. the land accurately. Gesture recognition is done after the features of the hands are extracted. Simple gestures are detected through the use of a set of predetermined rule based logic, whilst higher complexity gestures are detected through the use of a Convolutional Neural Network(CNN) model. This combination enables accuracy and enables the system to deal with the various forms of gestures effectively. After recognising the gesture it is mapped to certain medial control actions of play, pause, previous track, next track, and volume increase and volume decrease. These are carried out with the help of PyAutoGUI. It allows connecting with the media player directly. The above system has no requirement of a special hardware and is offline; thus it is economical and user friendly. This system is quite handy to old age users as well as differently challenged people.

B. Data Acquisition and Preprocessing

The data acquisition process begins with capturing a broad range of gesture samples using a standard webcam to build a reliable and diverse gesture dataset. Data from this system is collected using a webcam by recording different hand gestures. Each gesture is performed multiple times so that natural differences in hand size, shape are included in the dataset. After recording the gesture videos preprocessing is applied to improve the quality of the input frames.OpenCV is used for basic image processing tasks such as removing noise, resizing the frames and sharpening images when needed. MediaPipe is used to detect the hand in each frame and extract

21 important hand landmarks points. These landmarks represent finger joints and palm positions, which makes it easier for the system to understand the gesture without processing the full image. This reduces processing time and improves efficiency. The extracted landmarks are also normalized to handle changes in distance from the camera to improve performance data augmentation techniques such as flipping, scaling, rotating and adding small amounts of noise are applied. The measures assist the model to learn and perform well under the hands of various users. In general, preprocessing aids in maintaining normal and precise gesture recognition in actual use by the system.

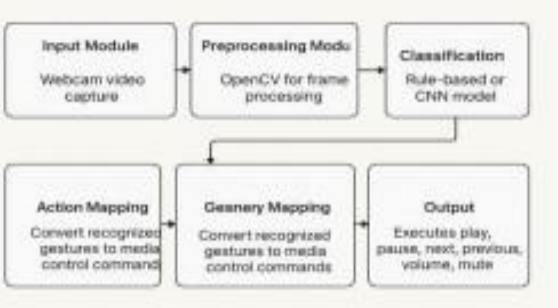
C. Feature Extraction and Gesture Embedding Generation

Hand gestures in this project are learned by finding some key points on the hand through MediaPipe followed by understanding them. It identifies the hand and provides 21 landmark points, in which the position of fingers and the palm are displayed.

These are sufficient points to explain the way the hand looks and moves thus they are used as the main features rather than using the complete image. As soon as the landmark points have been received the system checks what gesture should be carried out.

In simple gestures simple conditions are employed like the number offingers open or the distance between some points. The approach is fast and effective in gestures like play and pause. More complex gestures are represented with the aid of a CNN model. The model gets the patterns based on the land points and it generates a gesticulation of each gesture. This representation makes the system identify gestures even in the cases when the hand position is altered. The result or output of the model is then taken into consideration to determine the final gesture. Such a combination was adopted to ensure that the system remains quick and precise at the same time it also facilitates inclusion of new gestures in the gesture at a later age and also ensures that there are no problems in the real time interactions between the user and the system.

D. System Architecture



To achieve the gesture based media control the system architecture is developed by partitioning the whole process into various smaller modules, which collaborate with each other to produce the desired results. The input model employed the use of a web camera in taking live images of the user hand movement. These video frames normalization are applied to improve clarity and consistency. After preprocessing the feature extraction module uses MediaPipe to detect the hand and extract important landmark points. These landmarks are passed to the gesture recognition module, where the system identifies gestures. Simple gestures are recognized using rule based logic while more complex gestures are handled using a CNN model. The extracted features are then given to the gesture recognition module. For simple gestures the system uses straightforward rules based on finger positions. For more complex gestures, a CNN model is used to detect patterns in hand movement. After the gesture is identified, it is sent to the gesture mapping module, which decides the corresponding media action. Finally, the selected media command is executed using PyAutoGUI to control the media player.

A. Media Command Execution Workflow

When the system correctly detects a hand gesture, it connects that gesture to a specific media action. For example, showing an open palm can be used to play the media, a thumbs up gesture can skip to the next song, and a pinch gesture helps to increase or decrease the volume. The system performs these actions using PyAutoGUI by acting like normal keyboard controls. Because of this, it works smoothly with most media players and does not depend on any specific software. The action happens immediately, so the user feels a quick response. The system scans the similar gesture several consecutive video frames to prevent erroneous actions before executing the command. This helps ignore unethical or unintentional hand gestures. All operations are offline, and hence the system is rapid and no need of an internet connection to be safe. All this workflow simplifies the media control such that it becomes hands free and easy to operate even when it is not convenient to use physical control. The feedback is quick and safe as the system is offline.

IV. RESULT AND DISCUSSION

The system incorporates a safe system of entry to prevent people who are not authorized to use the gesture based media control application to enter the system. The authentication system is also made easy and easy to use so that even an individual with little technical knowledge can easily log in without any complication.

The essential features in the login interface are that it has the use of correct input validation, concealed passwords, and clear statements of errors to avoid wrong logins.

The measures also prevent the possibility of unauthorized or malicious access to the system.Secure session handling is provided to ensure the security of user credentials and the reliability of the session as long as the system is being used. Upon successful log in, the user automatically happens to be in the main control dashboard. Gesture recognition and media control functions can be initiated in this dashboard with ease. After logging in, the default screen displays a nice view of the current state of the system. It shows vital data like connection of webcam, gesture detection preparedness and presence of active gestures commands. This enables users to be in a position to know quickly whether the system is available to use or not.

Real-time updates are also displayed in the dashboard in terms of actions performed and gestures detected as well as performance of the system response. Such logs could assist the user to monitor the efficiency of the gesture recognition module in action.Users are able to modify system settings like sensitivity of gesture, speed of processing, and ability of detecting gestures depending on the environment or capability of his or her gadget. The dashboard is made to be user-friendly and simple, with visualization mechanisms such as status icons or progress bars to easily indicate whether the system is on or off or require re-calibration because of lighting or visibility problems.Generally, the dashboard is easy to operate and well-structured to access core controls easily as well as enhance user interfaces and monitoring the system.

Gesture Recognition and Capturing Process.

The user-friendliness of the system is provided by gesture capturing which is performed with the help of a live webcam feed. The video frame is repeatedly examined, and the existence of a hand is identified, and 21 significant landmarks are revealed via MediaPipe palm tracking technology.MediaPipe follows a multi-step procedure, which consists of palm recognition, hand area localization, landmarks localization.

partially covered. The system captures several frames per second with sufficient smooth and responsive gesture recognition with least delay. The system is made to be effective under various settings such as lighting variations, complexity of the background, and different skin tones and the hypotension of the hands. In order to make the detection more accurate, the interface offers visual feedback to make users modify their hand position, the distance between the camera and the light-sensitive objects, or the lighting in order to recognize command gestures: simple gestures are understood by a rule based logic, whereas more complex gestures undergo the classification by a CNN based classification model to learn spatial patterns of landmarks based on the data.

A. User Data Management and Gesture Performance Visualization

Its system will consist of an augmented user data area management as well as gesture performance visualisation area that present a holistic feel of the sufficiency of the gesture recognition process in real-time. The monitoring panel provides an understandable view of the system performance presenting such important analytics as the accuracy of gesture recognition, the level of confidence, the time and general success rates. The users can be informed of the functionality of the system in real time.

The number of frames that the system is processing, the average frequency of gestures detected and errors or false-detections are also recorded. This fact gives a better usage of the gestures recognition process that takes place indoors.

The panel can also present the panel graphicallowing the panel to show how precisely the system is monitoring the 21 hand landmarks. This enables the users to know whether they are properly using their hands. When some of the gestures suggest a low accuracy rate with a poor illumination case or any other position of the hand, one can recalibrate the system or re-train the model to provide a better performance.

By providing the performance information in transparent form, that is easily accessible, the monitoring module will instill confidence in the users and have the system remain reliable and malleable.

B. Execution of Media Controls and System output evaluation.

The system prompts the use of PyAutoGui to execute the relevant media action immediately a successful gesture has been identified. This enables the users to operate media easily with the help of smooth controls devoid of any physical input mechanism like a keyboard or mouse.

These confidence values allow the user to get an idea of how sure the system is of the gesture being detected with higher confidence the system can be in the output and reduce the chance of accidental or repeated commands.The system also determines that the gesture remains consistent across a series of frames before the system invokes any media control. Play, pause, next track, previous track, volume control, and mute operations have extremely low delays, creating a very responsive and easy-to-use experience.A real-time output list shows all actions taken to the user who can also ensure the system is responding appropriately to his/her gestures. Such systematic output assessment model leads to system stability, fewer errors and stable performance even within permanent gesture- based media control.

C. Discussions

The findings revealed that the proposed gesture-based media control system is very effective within a realtime environment with good accuracy, responsiveness and stability. The hand landmark detection offered by MediaPipe is significant to enhance the accuracy of tracking since it gives stable and reliable 3D coordinate reads of the hand even in an instance where the hand is moving modestly or executing intricate movements. The CNN- based classification model also enhances the ability to recognize better because it is more able to detect small differences in gestures which can not be easily identified by simple rule- based techniques.The user interface is user friendly and simple to operate thus enabling both the technical and the non technical users to use the system without having to undergo special training.

D. Output Screenshot

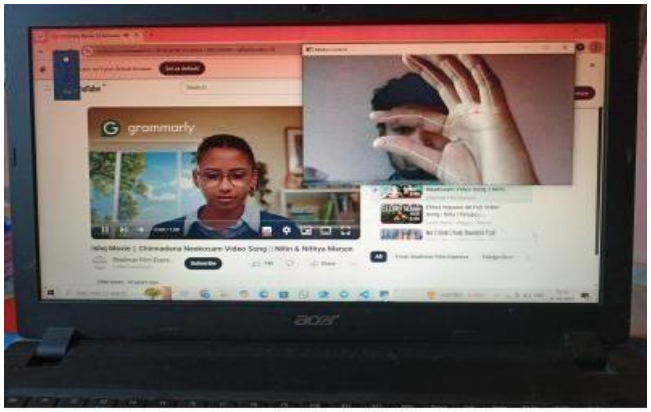


Figure 5.1: Increasing Volume

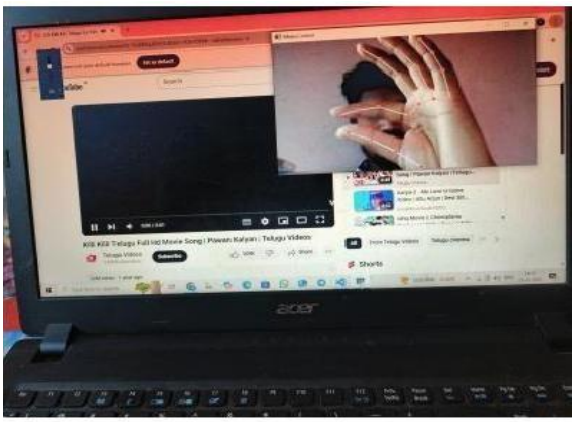


Figure 5.2: Decreasing Volume

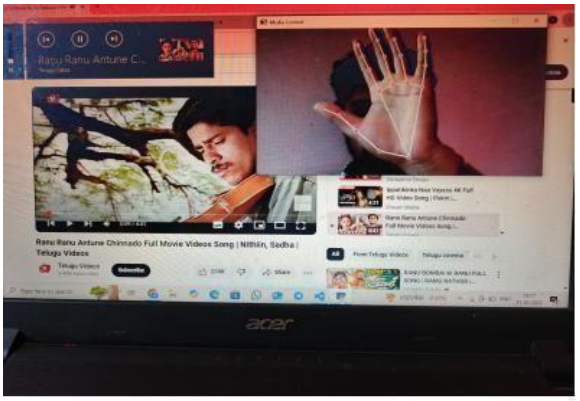


Figure 5.3: Play/Start

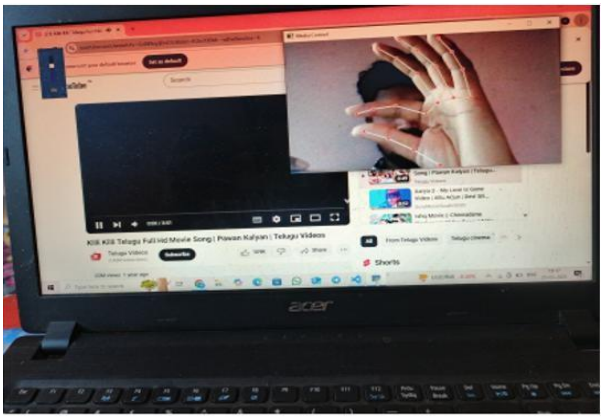


Figure 5.4: Decrease Volume.

V. CONCLUSION AND FUTURE SCOPE

The project being discussed is called Deep Learning Based Media Control with Gesture Recognition, which effectively illustrates the application of the artificial intelligence, computer and deep learning in developing a form of a human- computer interaction no-contact and user-friendly. A combination of rulebased techniques and Convolutional Neural Networks to perform gesture Vclassification was enabled through the use of MediaPipe hand landmark-detection, OpenCV video-processing, and the main objective of the real- time gesture-based media control system was reached.This way, the users can use simple hand signals to control the media on play, pause, next, previous, volume control, and mute as opposed to using the conventional input media like keyboards or mouse.

Its operation provides real-time performance by efficient preprocessing and enhanced gesture recognition, hence can be used on a daily basis. It is particularly handy to the elderly users, the differently-abled and an individual who needs to interact hands free as they multitask. The project is cost efficient, too, the project requires a conventional webcam and proper hardware only.

VI. REFERENCES

1 Google Research MediaPipe Hands: Real-Time Hand Tracking Solution Available at: https://google .github.io/mediapipe.
2 Bradski, G.(2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.
3 Ian Goodfellow, Youshua Bengio, and Aaron Couville. (2016). Deep Learning. MIT Press.
4 OpenCV Documentation. Open Source Computer Vision Library. Available at: https://docs.opencv.org.
5 hang, Z.(2012). Microsoft Kinect Sensor and Its Effect.
6 Krinzhevsky, A., Sutskever, I,. & Hinton, G.E.(ImageNet Classification with Deep Convolutional NeuralNetworks).
7 Chen, X., & Koskela, M.(2013). Online RGB-D Gesture Recognition with Extreme Learning Machines. Proceedings of the 15th ACM.
8 Saha, S., & Basu, S.(2020). Real-Time Hand Gesture Recognition Using MediaPipe and Tensorflow.
9 He, K., Zhang, X., Ren, S., & Sun, J.(2016). Deep Residual Learning for Image Recognition.
10 Wadhwa, N., et al, (2020). Real-Time 3D Hand Tracking Using MediaPipe and Deep Neural Networks, Google AI Blog.