IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

Video Streaming Platform with Automated Media Playback Control

Maddila Mohitha Shiva Sankari[1], Vasundhara Kandi[2], Sasidhar Karrothu[3], Sai Varun Karri[4], M. Beulah Rani[5]
Students, Department of Computer Science & Engineering[1-4]
Associate Professor, Department of Computer Science & Engineering[5]
Maharaj Vijayaram Gajapathi Raj College of Engineering (Autonomous), Vizianagaram, India
shivasankari2107@gmail.com, kandivasundhara2004@gmail.com, sasidharkarrothu@gmail.com, karrisaivarun@gmail.com, beulahrani@gmail.com

Abstract: *The project aims to develop a video streaming platform with automated media playback control, utilizing hand signals and voice commands . The main objectives include enabling users to control playback (play, pause, skip, volume) with gestures, also while combining voice commands for more intuitive control, and allowing customized gesture recording for personalized interactions and tailor controls to individual preferences. Expected outcomes include a significantly enhanced user experience with intuitive and accessible controls, innovative interaction methods driven by advanced technology leveraging machine learning algorithms and image processing techniques. The overall goal is to revolutionize media playback control, improving accessibility and user engagement in a competitive digital entertainment landscape.*

Keywords: Gesture Recognition, Voice Command Processing, Computer Vision, Machine Learning, Speech Recognition, Real-Time Processing, Flask, OpenCV, MediaPipe, VLC, YouTube Integration, Accessibility, Human-Computer Interaction

I. INTRODUCTION

During the age of consuming digital media, consumers increasingly demand more accessible and interactive interfaces with which they can engage with their entertainment. The conventional method of controlling the playback of media through remote controllers or touch controls has its limitation regarding convenience and usability. The project seeks to address these restrictions by offering an easier and free-hand mode of interaction, recognition of gestures and voice commands in real time, and innovations in machine learning and image processing boosting user experience while maintaining the platform in a technologically leading edge.

A. Identification of problem:

As there are fast-paced developments in digital entertainment, video streaming has also become a fundamental aspect of day-to-day life. Users communicate with media websites like YouTube, Netflix, and other streaming platforms using traditional input mechanisms, including remote controls, touchscreens, and keyboards. Although these inputs have worked so far, they have a few drawbacks:

- Persons who have physical disabilities or mobility impairments can find it challenging to operate typical input devices.
- Others, while doing other tasks (cooking, exercising, or working), can find it inconvenient to have to use their hands for manual control.
- Users must stop what they are doing to grab a remote or touch their device.
- It is not easy to navigate menus with remotes or keyboards.
- Conventional interfaces lack the ability to provide an interactive or immersive experience for controlling media.
- Also, the absence of multimodal inputs (e.g., gestures or voice) limits smooth interaction with media.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

To overcome these limitations, the blending of voice-based and gesture-based control mechanisms provides a contemporary, efficient, and accessible solution to media playback control. This is possible with a hands-free, natural, and intuitive way of interaction, enhancing user experience and accommodating accessibility requirements.

B. Key Objectives:

- Enables users to control video playback using predefined hand signals (e.g., raising one finger to mute/unmute, adjust volume, swiping to skip forward).
- Allows users to perform actions such as adjust speed, replay, rewind, and enter full-screen mode using voice instructions.
- Users can switch dynamically between voice and gesture commands, depending on convenience.
- Enhances user interaction by allowing multimodal input integration.
- The system will process both gesture and voice commands in real time, ensuring a smooth experience.
- A visual feedback mechanism will display the recognized gestures and commands, improving usability.
- The system will allow users to input a YouTube video URL and stream the video directly.
- By implementing machine learning algorithms, computer vision, and natural language processing, the system aims to enhance accessibility, interactivity, and user convenience in media playback control.

C. Significance of the system:

This paper primarily concerns applying machine learning algorithms and image processing methods, which will increase user interaction, accessibility, and overall convenience, making video playback interaction more efficient and immersive. The study of literature survey is presented in section III, Methodology is explained in section IV, section V covers the experimental results of the study, and section VI discusses the future study and Conclusion.

II. LITERATURE SURVEY

Human-computer interaction has come a long way, moving towards natural methods of input such as gestures and voice commands. Gesture recognition through image processing and machine learning approaches offers natural-looking control in numerous applications, such as media playback. Voice recognition with deep learning also makes hands-free use more effective.

Current work combines both modalities to enhance precision and user experience by using gestures for accuracy and voice for intricate commands. With progress, issues persist with real-time performance across changing conditions.

Automated Media Player with Hand Gesture by Priyadarshini Kannan, Sayak Bose, V. Joseph Raymond mentions that The current systems mainly concern the utilization of certain hand gestures or depth sensors for media player control, without offering flexibility for users to personalize their own gestures. The current systems do not incorporate facial expressions along with hand movements, which might give a richer and a friendly interface.

In a research by Al-Karawi et al. (2021), a hand gesture recognition system was implemented based on a deep convolutional neural network. The system was trained and evaluated on a hand gesture dataset captured with a webcam. The study results indicated that the deep learning model was highly accurate in recognizing hand gestures, proving the capability of the technology for application in human-computer interaction.

In Li et al.'s (2020) another study, a hand gesture-based control system was designed for the robotic arm. The system implemented a combination of recurrent neural networks and convolutional neural networks for identifying hand gestures and producing control signals for the robotic arm.The results of the study demonstrated the potential of HGR for controlling robotic systems in various settings, including manufacturing and healthcare.

In 2014, Swapnil D. Badgujar, “Hand Gesture Recognition System” proposed the system which recognizes the unknown input gestures by using hand tracking and extraction methods. This system is applied to recognize the single gesture. There is assumption of stationary background so that the system will have a smaller search region for tracking. This system only controls the mouse with the finger using it on the webcam.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

III. PROBLEM STATEMENT

The problem being addressed is the inherent limitations of traditional media playback controls, such as remote controls and touch interfaces, which can be inconvenient, restrictive, and inaccessible for many users. In response to the increasing digital media consumption, this project aims to develop a hands-free media control solution. By using real-time recognition of hand gestures and voice expressions this solution will provide a more intuitive and seamless way for users to interact with their content, enhancing overall accessibility and user engagement. The significance of this problem lies in the growing demand for more natural and intuitive ways to interact with digital media in an increasingly technology-driven world.

IV. REQUIREMENT GATHERING

A. Functional Requirements:

Media Playback Control

- The system should allow users to play, pause, rewind, and fast-forward videos.
- Users should be able to increase or decrease volume using hand gestures.
- Users should be able to mute and unmute the audio.

Gesture Recognition

- The system should detect hand gestures using a webcam.
- It should differentiate between various gestures (fist, 1-5 fingers, etc.).
- The detected gesture should be mapped to a media playback control.

Voice Command Processing

The system should process voice commands to:

- Adjust playback speed (increase/decrease speed).
- Rewind and replay videos.
- Toggle full-screen mode.
- The system should handle speech recognition errors gracefully.

YouTube Video Streaming

- The system should fetch and stream YouTube videos using a given URL.
- It should extract the direct video stream link using yt_dlp.

Real-Time Processing

- The system should process both voice and gestures simultaneously.
- It should provide instant feedback for recognized gestures and voice commands.

User Interface

- A visual display should show the webcam feed.
- The system should display a rectangular box for hand gesture detection.
- The detected action (e.g., “Volume Up”) should be displayed on the screen.

Error Handling & Logging

The system should handle errors like:

- Webcam connection failure.
- Invalid YouTube URLs.
- Speech recognition timeouts.
- Errors should be logged and displayed appropriately.

B. Non-Functional Requirements:

- Accuracy: High accuracy in recognizing gestures and voice commands.
- Scalability: Handle increasing numbers of users without performance degradation.
- Reliability: High uptime and minimal failures in gesture/voice recognition and media control.
- Accessibility: Ensure accessibility features, allowing users with disabilities to interact effectively.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

V. PROPOSED WORK

The process of developing the video streaming site with auto control of media playback is a sequential computational process incorporating artificial intelligence, computer vision, and speech processing methods. The approach includes some major stages ranging from system design to deployment in order to guarantee strong performance and user-friendliness.

The system design utilizes a client-server architecture with multimodal input processing. The frontend web application, developed using HTML/CSS/JavaScript, interacts with a Python Flask backend server that manages all processing units. The VLC media player is the central playback engine, while OpenCV and MediaPipe constitute the computer vision pipeline for real-time gesture recognition. This distributed design allows effective handling of simultaneous gesture and voice inputs while preserving low-latency response times.

For gesture recognition, we utilize a hybrid system integrating deep learning-based feature extraction and rule-based classification. Each video frame is processed by the MediaPipe Hands framework to locate and track 21 accurate hand landmarks, offering spatial locations of finger joints and palm orientation. The coordinates go through preprocessing via OpenCV, such as reducing noise and spatial normalization, prior to input to our gesture classification system. The categorization has deterministic rules determined by patterns of finger extension with certain combinations invoked for corresponding controls of media (e.g., one extended finger cycles mute on/off, five fingers play/pause).

Voice command processing leverages the Google Speech Recognition API with customized optimizations. The audio pipeline includes spectral subtraction noise reduction and endpoint detection to segment speech areas. Identified commands are keyword matched against a predefined list of media control verbs, with confidence thresholding to reduce false positives. The system features a command cooldown period to avoid duplicated executions from continuous speech input.

Data preprocessing also constitutes an essential part of our approach, especially for training the gesture recognition system. Our dataset consists of thousands of labeled hand gesture samples acquired under different lighting conditions and orientations of the hand. Each sample is cleaned to eliminate background noise, normalized to normalize spatial coordinates, and augmented by rotation and scaling to enhance model robustness. Feature selection targets relative finger positions and motion vectors between two consecutive frames.

VI. SYSTEM ARCHITECTURE

A. System Design

This section explains the flow on how the system works. The video streaming platform uses gesture and voice commands to control playback. The design consists of three main components:

1. Input Capture Layer

- Webcam (Gesture Input): Captures hand movements in real-time.
- Microphone (Voice Input): Records voice commands for processing.

2. Processing Layer

Gesture Recognition:

- Uses MediaPipe Hands to detect hand landmarks.
- Classifies gestures
    - One finger up → Mute/Unmute
    - Two fingers up → Volume Up
    - Three fingers up → Volume Down
    - Fist Gesture → Skip Forward
    - Five fingers open → Play/Pause

Voice Recognition:

- Converts speech to text using Google Speech Recognition.
- Matches keywords to actions.
    - "replay" → Plays from beginning

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

- "Increase speed" → increase speed by 0.5x
- “decrease speed” → decrease speed by 0.5x
- "Rewind" → Go back five seconds
- "Fullscreen" → Enter full-screen mode
- “normal screen” → exits full-screen mode

- If an unrecognized command is detected, an error message is displayed and no action will be performed.

3. Control & Playback Layer

- Flask Backend: Receives commands and sends them to the media player.
- VLC Player: Executes actions (play, pause, volume control).
- User Feedback: Shows on-screen confirmations for gestures/voice.

B. How It Works

- User opens the web interface and enters a YouTube link.
- The system starts capturing gestures and voice commands.
- Detected gestures/voice are processed and sent to VLC.
- The media player responds instantly (e.g., pauses when a hand gesture is recognized).

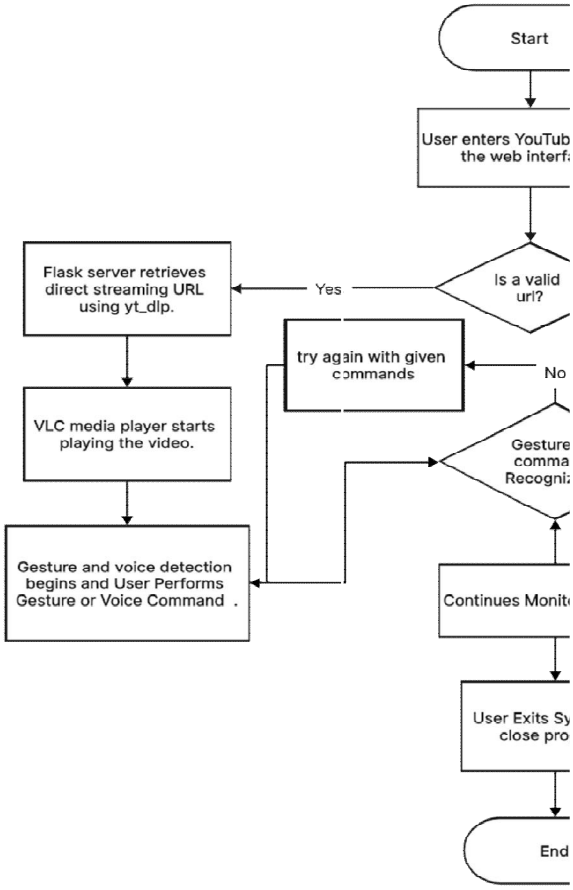

Fig1. flowchart of how system works

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

VII. RESULTS

The Implementation and testing of The Video Streaming Platform with Automated Media Playback Control has been done using gesture and voice recognition in real time. The system's efficacy was analyzed with respect to gesture recognition precision, voice command accuracy, command response time, and overall system responsiveness.

Video Streaming Module: To ensure proper extraction and playback of the YouTube video stream on VLC.Testing Approach:
Simulate responses from the YouTube API using predefined test URLs.
Check the format of the retrieved video URL.
Confirm VLC is initializing and playing the video.

Web Application:

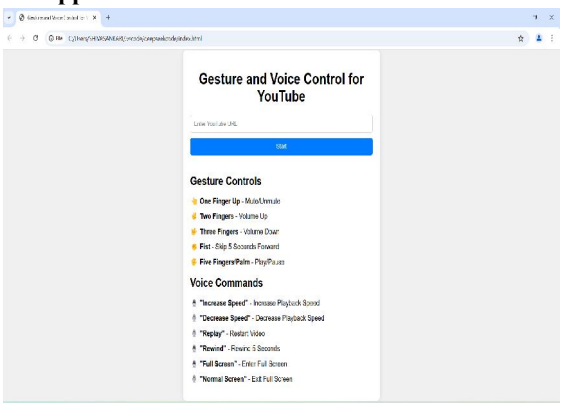

Fig2: web page

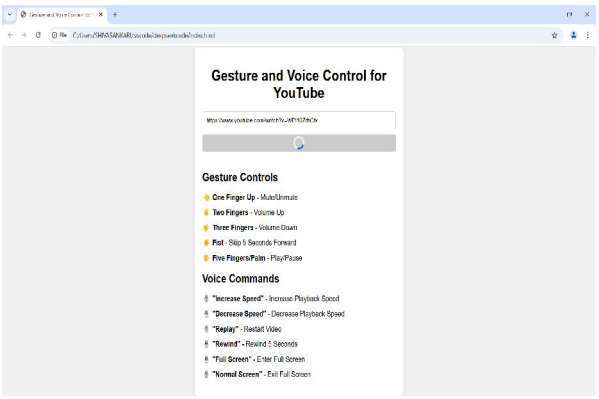

Fig3: Uploading video link in the url

Gesture Recognition Module: Ensuring hand gestures are detected accurately and mapped to correct playback actions.
Gesture Control Responses:

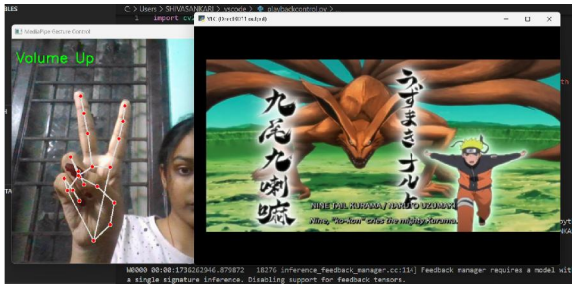

Fig4: Gesture Control - Volume up

Fig5: Gesture Control - Muted

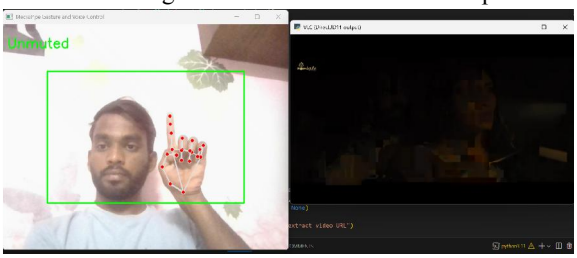

Fig6: Gesture Control - UnMuted

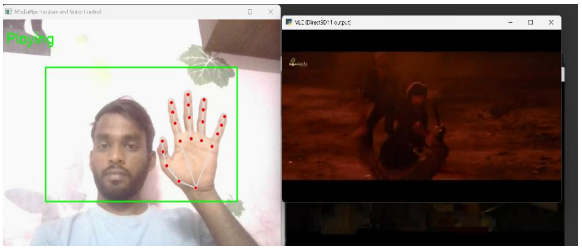

Fig7: Gesture Control - Play

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, April 2025 Impact Factor: 7.67

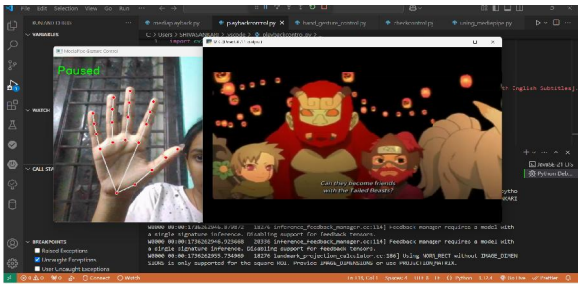

Fig8: Gesture Control - Paused

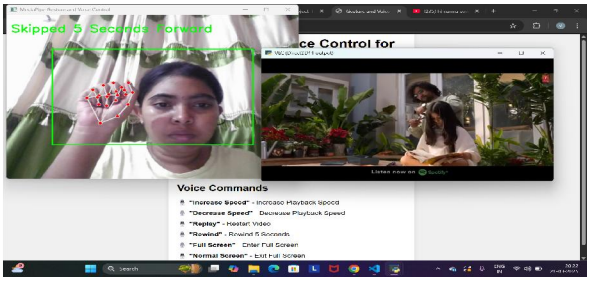

Fig9: Gesture Control - Skipped 5 Seconds Forward

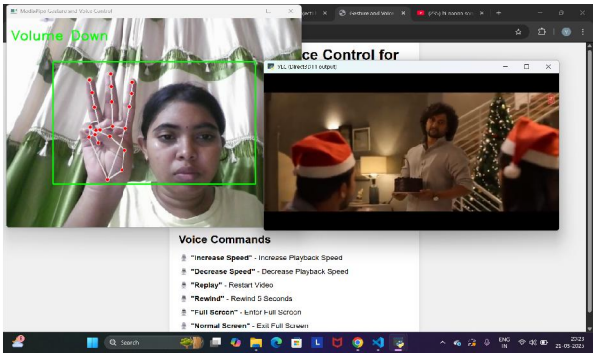

Fig10: Gesture Control - Volume Down

Voice Recognition Module : Ensuring spoken commands are correctly recognized and mapped to playback controls.

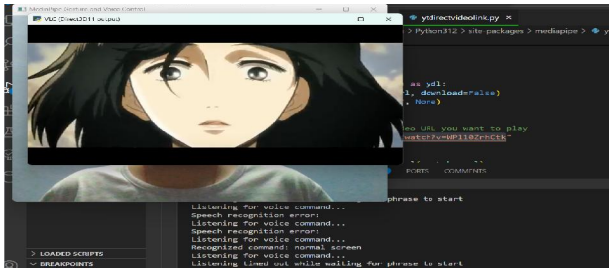

Fig11: Voice Recognition - Normal Screen

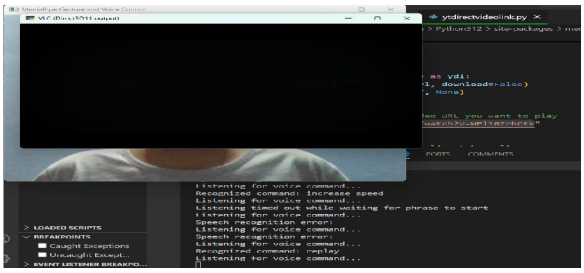

Fig12 : Voice Recognition - Replay

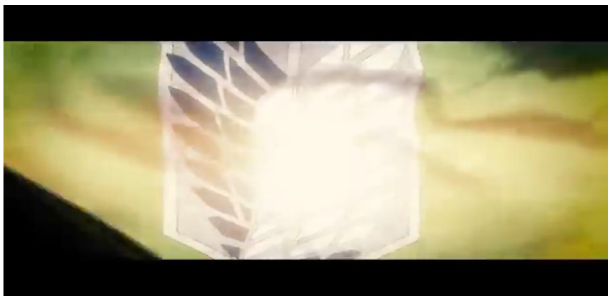

Fig13 : Voice Recognition - Full Screen

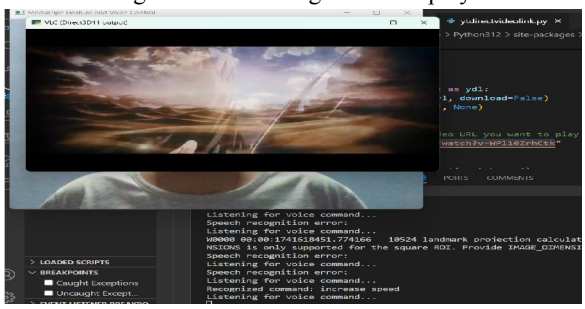

Fig14 : Voice Recognition - Increase Speed

DOI: 10.48175/IJARSCT-248440

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

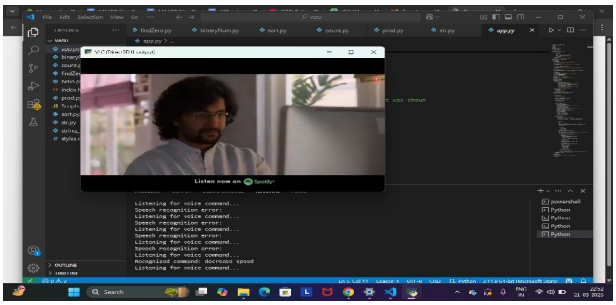

Fig15: Voice Recognition - Decrease Speed

VIII. CONCLUSION

This project effectively deploys a gesture and voice-controlled YouTube media player that improves accessibility and user experience. Through the combination of Flask, OpenCV, MediaPipe, VLC, and SpeechRecognition, users can navigate playback, volume and other controls without using their hands. The system exhibits real-time gesture recognition, seamless media integration, and visual feedback for user interaction.

It provides solid foundation for future works which may be extended to Support multi-language voice commands for wider accessibility, adding custom gesture recording to allow users to personalize controls and extend compatibility to smart TVs and IoT devices for hands-free entertainment control.

REFERENCES

[1] Kannan, P., Bose, S., Raymond, V. J., & International Research Journal of Engineering and Technology (IRJET). (2023). Automated Media Player using Hand Gesture. International Research Journal of Engineering and Technology (IRJET), 10, 1466.

[2] "A Real-Time Gesture Recognition System for Media Players using Convolutional Neural Networks" by J. Jung, H. Kim, and H. Park. This paper proposes a real-time gesture recognition system for controlling media players using a CNN.

[3]"Real-Time Hand Gesture Recognition for Controlling Media Players using Convolutional Neural Networks" by M. M. Asghar and F. Hussain. The authors proposed a real-time hand gesture recognition system for controlling media players using CNN and OpenCV.

[4] Rabiner, L. (1989). A Tutorial on Hidden Markov Models and selected models in Speech Recognition. Proceedings of the IEEE, 77(2), 257-284.

[5]Rao, A.S., & Selvan, S.E.L. (2013). Control Video with Head, Hand and Voice. International Journal of Computer Applications, 67(10), 41-44.

[6] Wen, Z. G., & Huang, K. (2010). RGB-D Sensor Based Personal Learning Gesture Using HMM Towards vision-based, in air, real excitable Human Computer Interactions.

[7] Preeti Sahu, Shashank Pathak, and Shailendra Singh - "Real-time hand gesture recognition for human computer interaction," International Journal of Advanced Research in Computer Science, vol. 7, no. 5, pp. 155-159 -2016.

[8] Aishwarya Shukla and Jyoti Verma, "Hand gesture recognition using SVM and SURF -" Procedia Computer Science, vol. 125, pp. 244-251 -2018.

[9] D. Yu and L. Deng, Automatic Speech Recognition: A Deep Learning Approach. New York NY, USA: Springer, 2014.