Smart Home Automation System

Shubham Sarkar
Department of Engineering
MIT-WPU College of Engineering
Pune, India

Akash Mohapatra
Department of Engineering
MIT-WPU College of Engineering
Pune, India

Neha Mendhekar
Department of Engineering
MIT-WPU College of Engineering
Pune, India

Chaitanya Dumbre
Department of Engineering
MIT-WPU College of Engineering
Pune, India

Dhananjay Bhagat
Department of Engineering
MIT-WPU College of Engineering
Pune, India

Seema Patil
Department of Engineering
MIT-WPU College of Engineering
Pune, India

***Abstract*—Smart home systems and Human-Computer Interaction (HCI) technologies have recently garnered attention for their potential to enhance daily life. This paper presents a multimodal interaction system addressing key challenges in creating seamless and intuitive interfaces by integrating voice commands, facial recognition, and gesture control in real-time. Our approach utilizes the realtimestt library for efficient speech-to-text conversion, Levenshtein distance for command correction, a Siamese network for personalized facial recognition, and an LSTM-based neural network for dynamic gesture recognition. By leveraging multithreading, this system achieves parallel processing, enabling simultaneous command processing, face verification, and gesture detection without compromising responsiveness. Experimental results demonstrate high accuracy in command recognition, low latency in processing, and reliable performance across diverse environments, showcasing the system's robustness and practicality. This work contributes a scalable solution for enhancing HCI in smart environments and highlights its potential for further advancements in multimodal interaction systems.**

***Index Terms*—Smart home systems, HCI, speech-to-text, face recognition, gesture recognition, real-time control**

I. Introduction

A. Problem Statement

Developing a smart home automation system that seamlessly integrates real-time voice command recognition, face identification, and gesture-based control to enhance user interaction and system responsiveness.

B. Objectives

- **Integrate Multimodal Inputs:** Develop a system that combines real-time speech-to-text, face recognition, and gesture detection for seamless control in smart home automation.
- **Enhance System Accuracy and Responsiveness:** Optimize voice command recognition, face identification, and gesture-based control to ensure high accuracy and low latency in dynamic home environments.
- **Enable Hands-Free Automation:** Create a user-friendly interface that allows intuitive, hands-free control of smart home devices, improving convenience and efficiency for users.

C. Background

The rise of smart home automation demands more intuitive and natural interfaces beyond traditional remote controls or apps. Multimodal systems incorporating voice commands, facial recognition, and gesture control offer hands-free, personalized interaction, but face challenges in accuracy and responsiveness. This project addresses these challenges by integrating these technologies into a seamless and efficient smart home control system.

D. Importance

This project is important as it advances smart home automation by providing a more natural, intuitive, and hands-free control system. It improves user convenience, enhances accessibility, and addresses the challenges of accuracy and responsiveness in multimodal human-computer interaction, leading to a better smart home experience.

II. Literature Review

A. Face Recognition using Siamese Network [1]

Facial recognition systems have advanced in security and attendance tracking, but traditional methods face challenges with pose, aging, and lighting. Recent approaches use deep learning, especially Convolutional Neural Networks (CNNs) and Siamese Networks, to improve accuracy. CNNs extract keypoints for identification via K-Nearest Neighbors (KNN), while techniques like Particle Swarm Optimization (PSO) enhance performance. Innovations such as triplet loss in Siamese Networks further refine extraction. The technology's use in healthcare and e-voting underscores its importance, highlighting the need for robust, low-intervention models. This survey reviews current methodologies and their implications for future research.

B. Design of Home Appliance Control System in Smart Home based on WiFi IoT [2]

WiFi-enabled microcontrollers like the ESP8266 and ESP32 are popular for controlling devices such as LEDs and motors via HTTP requests. Acting as servers or clients, they allow real-time control from mobile apps or browsers. Studies

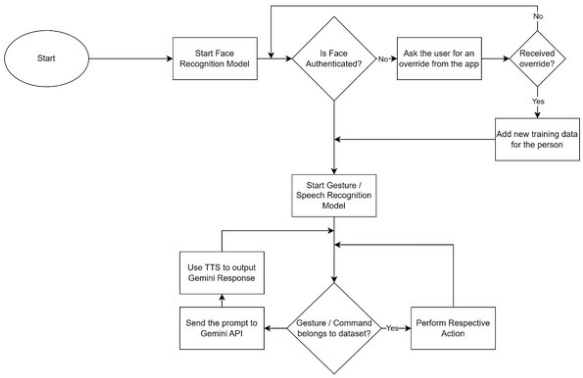

Fig. 1. Flowchart

highlight their low cost and efficiency for IoT-based home automation. Relays are often used alongside ESP modules to control high-power devices like motors. HTTP-based control integrates well with cloud services, enabling remote operation and automation.

III. Methodology

A. Face Recognition [3]

- The MTCNN model from Facenet-Pytorch is used to detect all the faces in the image.
- These faces are processed through InceptionResnetV1 to generate embeddings for each face.
- These embeddings are then compared to already learnt embeddings using Euclidean Distance as part of the Siamese Network process.
- If the distance lies within the threshold, the closest one is used to identify the person; otherwise, the face is considered 'Unknown' and access is not granted.

B. Speech Recognition

- Using the Real-time STT Python library, speech recognition is performed to transcribe user speech.
- The phonetic representation of the recognized text is obtained using the Soundex algorithm.
- This Soundex value is compared to those of known voice commands, and the one with the best match is chosen.
- In case no match is found, Levenshtein distance is calculated between the recognized text and known voice commands, and the one with the least distance to the recognized text is chosen as the actual text if within a threshold.
- If the user used a voice command, the respective API call is sent to the ESP for actuation; otherwise, the prompt is sent to the Gemini API to get a suitable response.

C. Gesture Recognition [5]

- When a voice command activates the gesture detection system, OpenCV starts recording video.
- The recorded frames are fed into MediaPipe (and optionally OpenPose) for feature extraction.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) \ if \ \min(i,j) = 0, otherwise \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} \end{cases}$$

Fig. 2. Levenshtein Distance

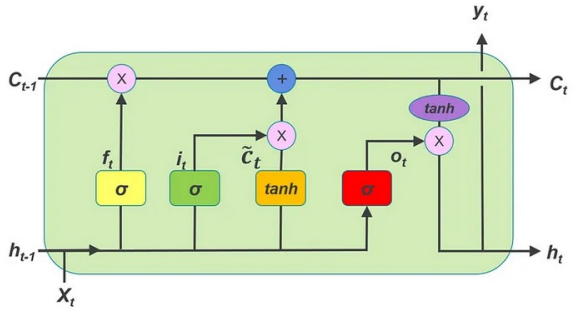

Fig. 3. LSTM

- Extracted features include hand landmarks, angles between the wrist and index finger, and other important metrics.
- The extracted features, paired with their labels, are used to train a machine learning model.
- The model is a combination of a neural network and an LSTM, designed to recognize dynamic gestures over time.
- After training, the system processes live video input from the camera in real time, extracting features and making predictions about gestures.
- When a specific gesture is recognized, a signal is sent to an ESP board to control the associated hardware component.

D. Actuation [10]

- When the ESP module receives an API request triggered by a recognized user command, it processes the instructions and translates them into actions for the connected devices, such as fans and lights, ensuring proper actuation.
- **Fan control**: The ESP modulates the fan's speed and direction using a float value. This value controls the velocity, adjusting the motor driver to increase or decrease speed, and can also reverse the fan's direction if needed.
- **Light control**: The ESP uses a binary input to switch the lights on or off. If dimming is supported, it adjusts the brightness based on the float value, ranging from completely off to full brightness.
- After executing the command, the ESP provides feedback to the system or user, confirming the action and keeping the device status in sync with the issued commands.

IV. Results

A. Face Recognition [3]

- Initial testing with CNN for face recognition required more image data for accurate recognition and security.
- The Siamese Network method, coupled with feature extraction using InceptionResNet, is able to provide accurate face recognition with very little data.

B. Speech Recognition

- Initially, the Real-time STT implemented was not completely accurate with its speech recognition and often misinterpreted phrases. For example, "Lights On" was sometimes recognized as "Likes On."
- To fix these problems, the Soundex algorithm was used to find the most appropriate conversion to a command if the distance to it was within the threshold.

C. ESP Code [10]

- Initial code for the ESP8266 module had delays and inefficiencies when managing multiple device commands.
- To address this, the code was optimized to handle commands simultaneously, ensuring smoother operation.
- The fan's speed and direction were dynamically controlled using input values, while the lights were switched on/off.
- Real-time feedback was added to confirm the execution of commands, keeping the system synchronized.
- After optimization, the ESP code ensured faster and more reliable device control across the system.

D. Gesture Recognition [5]

- Simple gestures were recognized accurately, but complex and dynamic gestures faced classification challenges due to overlapping movements and noise.
- Using MediaPipe and OpenPose for feature extraction improved accuracy, especially for gestures involving intricate hand and finger movements.
- The LSTM network effectively managed dynamic gestures by recognizing sequential movements over time.
- The system accurately detected gestures in real time, allowing smooth interaction with ESP-controlled devices.
- Gesture-based commands successfully triggered actions like fan speed adjustment and light control, confirmed by real-time feedback from the ESP.
- The system demonstrated reliable recognition under varied lighting and hand positions, ensuring robust operation.

V. Discussion

A. Strengths

- Face recognition is easy to train using very few images.
- The modular design ensures that all models can be changed easily.
- ESP actuation is handled separately using an API, so minimal changes are needed to modify the functioning.

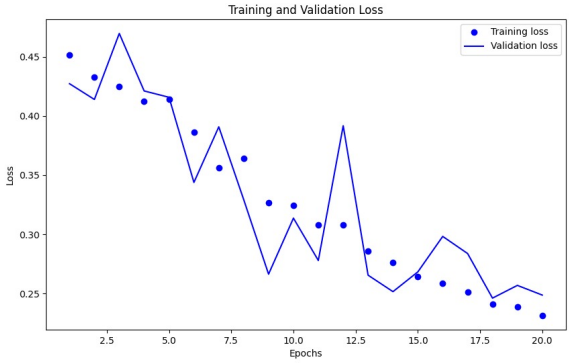

Fig. 4. Training and Validation Loss

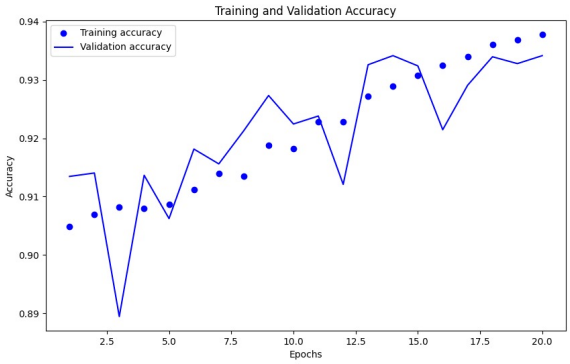

Fig. 5. Training and Validation Accuracy

B. Future Work

- Research can be conducted to add speaker identification for enhanced security, which would not require the user to stand next to the camera.
- Motion sensors can be used to optimize when the Real-time STT starts and stops, reducing energy consumption.

VI. Conclusion

Preliminary tests of the smart home automation system have demonstrated promising results, showing that the system accurately identifies and verifies authorized users within a predefined range. The face recognition module effectively distinguishes between authorized and unauthorized users, enhancing the security of the system by preventing unapproved access. This ensures that only registered users are able to interact with the home appliances, adding an extra layer of protection.

In terms of responsiveness, both the voice and gesture recognition modules perform efficiently, registering commands with minimal latency. Voice commands, processed using natural language processing (NLP), allow for smooth control over appliances like lights and fans, while the gesture recognition system provides an alternative mode of interaction. Together,

these input methods create a highly intuitive and versatile user experience.

The system's reliance on wireless IoT communication has also proven to be reliable, enabling fast and accurate transmission of commands to the connected devices. The appliances respond almost instantly once the commands are issued, reinforcing the system's practicality. The combination of these technologies results in a smart home automation system that not only ensures a high degree of security but also provides a seamless, user-friendly interface for managing home environments.

VII. References

References

[1] M. Heidari and K. Fouladi-Ghaleh, "Using Siamese Networks with Transfer Learning for Face Recognition on Small-Samples Datasets," 2020 International Conference on Machine Vision and Image Processing (MVIP), Iran, 2020, pp. 1-4, doi: 10.1109/MVIP49855.2020.9116915.

[2] Z. Xiao, D. Liu, D. Cao and X. Wang, "Design of Home Appliance Control System in Smart Home based on WiFi IoT," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2018, pp. 765-770, doi: 10.1109/IAEAC.2018.8577217.

[3] H. Wu, Z. Xu, J. Zhang, W. Yan, and X. Ma, "Face recognition based on convolution siamese networks," Oct. 2017, doi: https://doi.org/10.1109/cisp-bmei.2017.8302003.

[4] S. Qiao, Y. Wang and J. Li, "Real-time human gesture grading based on OpenPose," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 2017, pp. 1-6, doi: 10.1109/CISP-BMEI.2017.8301910.

[5] Y. Wu, B. Zheng and Y. Zhao, "Dynamic Gesture Recognition Based on LSTM-CNN," 2018 Chinese Automation Congress (CAC), Xi'an, China, 2018, pp. 2446-2450, doi: 10.1109/CAC.2018.8623035.

[6] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021)

[7] Hochreiter, Sepp and Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[8] Lugaresi, Camillo, et al. "Mediapipe: A framework for building perception pipelines." arXiv preprint arXiv:1906.08172 (2019).

[9] Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).

[10] Macheso, Paul, et al. "Design of standalone asynchronous ESP32 webserver for temperature and humidity monitoring." 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). Vol. 1. IEEE, 2021.