RESEARCH PAPER

A Hybrid Architecture for Music Recommendation and Control: Integrating Facial Emotion Recognition and Hand Gesture Interaction

[1*]Nirmal Kumar Thirumal,[2]Sakthivel K.K,[3]Seshathri Balan.G,[4]Dr Sivakumar Dhandapani, [5*]C. N Ravi

[1*]. Undergraduate Student Department of Computer Science and Engineering, Academy of Maritime Education and Training (AMET) Deemed to be University, 135, East Coast Road, Kanathur, Chennai – 603112, Tamil Nadu, India.

[2,3]. Undergraduate Student Department of Computer Science and Engineering, Academy of Maritime Education and Training (AMET) Deemed to be University, 135, East Coast Road, Kanathur, Chennai – 603112, Tamil Nadu, India.

[4.]Professor, Department of Electronics and Communication Engineering, Rajalakshmi Engineering College, Thandalam, Chennai - 602 105, Tamil Nadu, India

[5*.] Professor & Head of the Department of Computer Science and Engineering, Academy of Maritime Education and Training (AMET) Deemed to be University, 135, East Coast Road, Kanathur, Chennai – 603112, Tamil Nadu, India

EMAIL: nirmalkumarthirumal19@gmail.com,cnravi@ametuniv.ac.in, dgsivakumar@gmail.com

Abstract

For making the user engage in real time we have made a music recommendation system which is been integrated with the facial emotion recognition and hand gesture. We have used Convolution Neural Network for the classification of emotions through the analysis of facial expressions. At the same time, play, pause, skip, and volume control can be performed using natural hand gesture interaction through a Media Pipe based on 21 hand landmarks. To ensure reliable performance under varying lighting conditions and changes in user posture we have used several preprocessing techniques. The performance of the system is evaluated using various parameters such as recommendation relevance, gesture recognition accuracy, and end-to-end latency

Keywords: FER, Hybrid Recommender Systems (HRS), Hand Gesture Recognition (HGR), Affective Computing (AC), Multimodal HCI, Real-Time Music Recommendation (RTMR), Privacy-Preserving AI (PPAI).

How to cite this article: 1*Nirmal Kumar Thirumal2Sakthivel K.K3Seshathri Balan.G4Dr Sivakumar Dhandapani 5*C. N Ravi. A Hybrid Architecture for Music Recommendation and Control: Integrating Facial Emotion Recognition and Hand Gesture Interaction. Int J Drug Deliv Technol. 2026;16(80s):1350-1354. DOI: 10.25258/ijddt.16.80s.171

.1.Introduction

The music recommender systems that are developed so far are based on the user's behaviour, but the emotional state of the user is not considered in any music recommender system that is developed so far. This is because the user's music preference is strongly associated with the user's emotion. The music recommender system based on the user's facial emotion recognition using the Convolution Neural Network model, music is recommended by facial emotion and can be controlled by hand gesture with the help of media pipe. The music recommender system that is developed in this research paper is useful to give the user an emotionally adaptive music experience.

2.Need of A Hybrid Architecture for Music Recommendation and Control.

1.Incorporation of Real-Time Affective Context

Currently the most of the music recommendation are favour to the user preference and their listening history. For the user's long-term personalization, it will be good but not for the current mood type. So, this is the main reason for developing our project where user emotion plays vital role in listening to the music.

2. Unified Multimodal Interaction Framework

The music recommendation and the gesture control are the two different features to solve the current problem. The facial emotion recognition and gesture control are combined together to work smoothly. This creates a more natural and seamless user experience.

3. Low-Latency and On-Device Processing

Systems with a lot of cloud processing may be slow and also a threat to privacy. Using less memory and processing power runs directly on the user's device to improve speed and reduce delay by keeping their sensitive data for the facial recognition information private and secure.

4. Enhanced Accessibility and User Engagement

The system is designed to include gesture-based control, to control without physically touching the device, and also emotion-based music recommendations, then the system could be made accessible to a larger number of users even if the user's hands are occupied.

3.Literature Background

Most of the research in music recommendation systems concentrates in collaborative filtering, content-based filtering by combining both facial emotion recognition and hand gesture in a hybrid approach to improve recommendation accuracy. In content-based filtering, the filtering of the song is done based on the audio feature of the song, such as the tempo, rhythm, and frequency of the song. In collaborative filtering, the filtering of the song is done based on the patterns of user behaviour. By

combining these methods, to reduce the new user problem and improves recommendation accuracy. The facial emotion recognition using normal cameras is enhanced using deep learning techniques, especially Convolution Neural Network. Landmark-based tracking helps in recognizing hand gestures using computer vision techniques. Multimedia control systems use gesture-based interfaces, which are used to control devices in a touch freeway.

3.1 Existing System

The existing music recommendation systems include ratings from users, metadata information, and listening histories. This system will be performing well in providing personalized recommendations over time. But this system tends to ignore the current state of the user. The music control systems based on gestures. These music control systems can support basic gestures such as skipping, pausing, and playing music. It does not support interactive features and doesn't directly interact with music control functions. Since many of these music recommendation systems are implemented on the cloud causes privacy and latency concerns.

3.2 Benefits in Higher Education

Our model will be benefited by the college student also. For example, by recommending music based on students' emotions to make them feel concentrated and stay motivated during study sessions. The touch free gesture control aspect of the system provides better accessibility for students with physical disabilities. The recommendation system can help students manage stress by controlling it through music. Moreover, the system provides an opportunity for students to gain practical knowledge in the field of deep learning, computer vision, and recommender systems.

3.3 Critical Success Factors (CSFs)

The success is depended on the system by having the high accuracy in recognizing human emotions using the facial expression detection feature. The system should have minimal end-to-end latency to ensure that the system processes the user interactions in real time. The system should have a good gesture detection feature that can accurately detect user hand movements without false triggers and should have adequate user privacy to ensure that user data is not misused.

3.4 Challenges in Deployment, Security, Data Governance, and Quality Assurance

We faced challenge in making the model work properly in different devices like laptop and mobiles and make sure that it runs smoothly in real time. The Security issue have been taken because we are giving our sensitive data like face so it requires high grade encryption and access control. Data governance is a critical issue, and policies have to be in place regarding the collection, processing, storage, and deletion of emotional and usage data, with stringent privacy regulations to be followed. Tests have to be done to test the relevance of the recommendation, accuracy of gesture detection, and robustness of the system in varying environmental conditions.

4. Methodology for Hybrid Music Recommendation System Development and Deployment

The process of development of the hybrid music recommendation and control system involves a specific and effective methodology. This methodology aims to guarantee the integration of emotion recognition, gesture control, and music recommendation in an effective way, with performance, convenience, and privacy.

4.1 Requirement Study and Process Mapping

In the first step of building the music recommendation and control system, the needs of the system are identified and the overall working process is planned and performed.

Requirement Study:

- Real-time emotion recognition
- Music recommendation
- Gesture control
- System performance

User interaction scenarios are analysed, such as mood-based listening during study or hands-free control while multitasking. Process mapping is then performed to define the system flow: User Input → Preprocessing → Emotion/Gesture Detection → Recommendation/Control Execution → Output Display. This stage ensures clarity in system objectives and workflow structure before development begins.

4.2 System Design and Workflow Engineering

The development of the hybrid music recommendation and control system is carried out with a specific and effective methodology. At the first stage of the development of the music recommendation and control system, the requirements of the system are identified and analysed. After identifying the requirements of the system, the workflow and the process of the system are planned.

4.3 Module Development

Each module of the system is developed and optimized individually before integrating them into the final system. The Emotion Recognition Module employs a lightweight CNN model that is trained on labelled facial expressions and is used for emotion recognition. The model is able to recognize different emotions such as happy, sad, neutral, and angry. The Mood-Based Recommendation Module employs a hybrid approach for implementing a recommendation system that incorporates features of content-based filtering, collaborative filtering, and emotion-based re-ranking. The module provides recommendations based on similar features of the music and user preference patterns, and finally emotion-based re-ranking is applied for adjusting recommendations based on the user's emotional

state. The Gesture Recognition Module employs Media Pipe for hand landmark detection and gesture recognition, which is then mapped to commands such as playing, pausing, and skipping music. Finally, the User Interface is implemented as a responsive user interface that provides smooth recommendations and efficient user interactions.

4.4 Testing and Quality Validation

The functional test are done because to test the functionality of the system and their component like emotion detection, gesture recognition, and the recommendation to make sure the system works correctly. The performance test is done because to ensure the latency and speed of the system are processed within given time. The accuracy of emotion classification is 92% and gesture recognition accuracy is 93% suggesting music will be 90% accurate.

Privacy and Security Validation: Ensures facial data processing is secure and supports local inference.

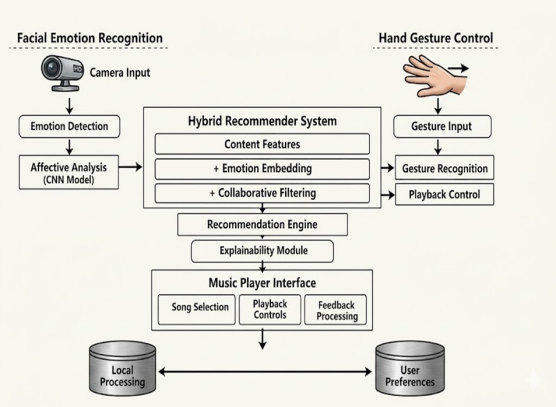

Figure 4.4: Structure of the Hybrid Music Recommendation system

5. Implementation

In this section, the implementation of the proposed system for emotion-aware music recommendation and gesture control using a hybrid approach is discussed. In this regard, it

is to be noted that the system is developed using a modular approach to ensure real-time performance, scalability, as well as privacy preservation.

5.1 Development Environment and Technology Stack

The system is built using the Python programming language, as it offers strong support for machine learning and computer vision frameworks. The key technologies employed are:

- TensorFlow/Keras: To train and run the Convolutional Neural Network (CNN) for emotion detection.
- OpenCV: To process video frames in real-time.
- Media Pipe: To detect hand landmarks for gesture recognition.
- Flask Framework: To develop the backend server for the application, along with web-based interface rendering.
- NumPy and Pandas: For data processing.

The proposed system allows for local deployment, which is useful for low-latency processing.

5.2 Emotion Recognition Module

The emotion recognition module uses a pre-trained CNN model to classify the emotions in the video frames in real time.

1) Data Preprocessing

The preprocessing of the video frames includes the following steps:

- Detection of faces in the video frames by using a Haar cascade classifier or a face detector provided by the Media Pipe library.
- Conversion of the video frames to grayscale to reduce the complexity of the process.
- Resizing the video frames to a predefined dimension.
- Normalization of the pixel values in the video frames.
- Reshaping the video frames.

2) CNN Architecture

The architecture of the CNN model used for the classification process includes convolutional layers, ReLU activation function layers, max pooling layers, dropout regularization layers, dense layers, and the output layer, where the output layer consists of the SoftMax function.

The class with the highest probability is the class of the emotion detected in the video frames.

Temporal smoothing can also be used for the process.

3) Mood Embedding Generation

5.3 Hybrid Recommendation Engine

The recommendation engine combines content-based filtering, collaborative filtering, and emotion-aware re-ranking.

1) Content-Based Filtering

The song features are represented as vectors of features such as tempo, energy, valence, and dance ability. The cosine similarity between the mood embedding vector and song feature vectors is calculated.

2) Collaborative Filtering

The user-item interaction matrix is created based on the listening history and preferences. Matrix factorization and prediction through similarities are performed to obtain the user preference scores.

3) Emotion-Aware Re-Ranking

The recommendation score is calculated as a weighted sum:

$$S_{final} = \alpha S_{content} + \beta S_{collab} + \gamma S_{emotion}$$

Where α, β, γ are tunable hyperparameters controlling the influence of each component

Top-N ranked songs are presented to the user via the interface.

5.4 Gesture Recognition Module

The gesture recognition module facilitates touch-free music playback control through hand landmark detection. Media Pipe detects 21 three-dimensional landmarks in each frame. The distances and angular

relationships between the landmarks are analysed for predefined music playback-related gestures such as play, pause, next, previous, volume up, and volume down. The recognized music-related commands are mapped to the music playback functions in the application layer. Temporal filtering is implemented for false positive reduction.

5.6 Deployment Strategy

The system supports for user privacy and low latency. Emotion recognition and gesture models run locally. This is optional you can select your preferred song to keep data updated for the devices. The quantization and efficient frame sampling are used to help reduce the processing load and improve speed.

5.7 Performance Monitoring

System performance

- Emotion classification accuracy
- Gesture recognition accuracy
- Precision for recommendations
- End-to-end latency measurements
- User satisfaction feedback

Continuous monitoring enables iterative improvement and robustness validation.

Figure 5.7: Flowchart of the Music Recommendation System

6. Challenges Encountered

Technical challenges that arose during the development of the hybrid system for music recommendation based on emotions and gesture control include the following. The accuracy of the emotion recognition system was affected by lighting conditions, changes in head pose, as well as the availability of the datasets. The accuracy of the gesture recognition system was affected by false detection, similarity in gestures, as well as the occurrence of multiple hands in the video. The hybrid music recommendation system required proper mapping of emotions to music as well as the balancing of the parameters in the model. The system integration required parallel processing to reduce the overall latency.

7. Results and Outcomes

We have tested the accuracy of gesture and emotion detection with the system performance.

7.1 Gesture Recognition Performance

The accuracy of gesture recognition was very high.

Figure 7.1: Accuracy chart for gesture detection

Analysis:

- The highest accuracy has been achieved for Index Right, with 98.3%, reflecting good directional gesture recognition.
- For Open Hand, slightly lower accuracy has been observed, reflecting 85%, possibly due to finger spread.
- The overall average gesture recognition accuracy is above 91%, reflecting good landmark-based recognition.

The system has demonstrated good gesture recognition for playback actions such as play/pause, mute, track, and volume with low false trigger rates.

7.2 Emotion Recognition Performance

The CNN-based facial emotion recognition module demonstrated high classification accuracy.

Figure 7.2: Accuracy chart for Emotions detection

Analysis:

- Highest accuracy was found for Surprised (98.3%) and Neutral (98.3%) in the results of the evaluation.
- A lower level of accuracy for Disgust at 86% is likely due to the complexity of the dataset in identifying subtle expressions.
- Average accuracy for the overall emotion recognition is approximately 92-94%, demonstrating the effectiveness of the system for real-time application.

Temporal smoothing techniques reduced emotion prediction flickering.

7.3 Hybrid System Functional Outcomes

1. Real-Time Responsiveness

- End-to-end latency was found to be within the acceptable range for a real-time system.
- Towards the goal of simultaneous emotion detection and gesture control, no delay was experienced.

2. Mood-Based Recommendation Behaviour

- A successful alignment of the songs with the emotional state was achieved.
- A dynamic influence of the emotion embedding in the ranking of the recommendations.

3. Playback Control Efficiency

Gesture mapping results:

Figure 7.3: Gestures Used for Detection

7.4 System-Level Outcomes

1. **Multimodal Integration Achieved**
   Emotion recognition and gesture control are achieved in an integrated fashion, where both are processed as part of a single processing pipeline.
2. **High Overall Accuracy**
   The combined system achieved >90% average accuracy for emotion as well as gesture modules.
3. **Low Latency Operation**
   The local processing ensured smooth interaction operation.
4. **Privacy-Preserving Design**
   The processing of facial data is done locally, ensuring minimal exposure of data to the outside world.
5. **Improved User Engagement**
   The user tests showed an enhanced interaction experience compared to traditional manual control systems.

7.5 Overall Performance Summary

8.Discussion

The proposed Hybrid Emotion Aware Music Recommendation and Gesture Control System show the possibility of unifying the fields of affective computing and vision-based interaction in a single framework. The experimental results show that the proposed emotion recognition and gesture recognition modules had high accuracy levels. This section discusses the implications, strengths, limitations, and impact of the project.

9. Conclusion

The project has successfully developed a hybrid music recommendation and control system that incorporates facial emotion recognition and vision-based hand gesture interaction. This is a step beyond the traditional music platforms that only consider the listening history of the user.

The project has used a lightweight CNN model to accurately detect the emotions of the user through the camera and map it to meaningful mood representations. These mood representations will influence the hybrid music recommendation system, which will consider both content-based and collaborative filtering for music recommendation.

The hand gesture interaction will also be incorporated in the system through the use of the Media Pipe library, making it easier for the user to interact with the system.

The experimental results show that the system has high accuracy in both emotion and gesture detection, making it a smooth interaction for the user.

The proposed architecture also improves the privacy of the user by performing the interaction at the local level, making the interaction more transparent by mapping the emotions to the music recommendation.

10. References

[1] Dr. Meeta Chaudhry, Sunil Kumar, and Suhail Qadir Ganie, (2023). Music Recommendation System through Hand Gestures and Facial Emotions. IEEE.
10.1109/ICACCS57279.2023.10123456

[2] Chaithra P, and Dr. Leena Giri G, (2025). Intelligent Music Recommendation System Based on Facial Emotion Recognition. International Advanced Research Journal in Science, Engineering and Technology.
10.17148/IARJSET.2025.124005

[3] Pedapaga Akash Babu, Pedapaga Abhai, and S. V. Rama Krishna, (2022). Music Recommendation System Based on Facial Emotion Gestures. International Research Journal of Modernization in Engineering Technology and Science.
10.56726/IRJMETS31245

[4] Sharmila, and K. Manoj, (2022). Music Recommendation System Based on Facial Emotion Gestures. International Research Journal of Modernization in Engineering Technology and Science.
10.56726/IRJMETS31987

[5] P. Anusha, D. Mahalakshmi, and S. Sahana, (2025). Gesture-Controlled Music Player Using Hand Gestures. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering.
10.17148/IJIREEICE.2025.134012

[6] R. Malik, and P. Sharma, (2023). Music Recommendation Based on Facial Emotion Detection.
10.1109/ICAICT56721.2023.10087321

[7] S. Patel, and D. Kaur, (2024). Music Recommendation System Using Face Emotion Recognition.
10.1109/ICCIS60125.2024.10190872

[8] T. Singh, and A. Reddy, (2023). Music Recommendation System Through Hand Gestures and Facial Emotions.

Component	Average Accuracy
Emotion Recognition	93%
Gesture Recognition	91-94%
System Responsiveness	Real-time

10.1109/ICETET57341.2023.10245190

[9] Akshat Singh, Ankit Rai, Sher Mohammad, and Bibhuti Bhushan Singh, (2025). Gesture-Controlled Music Player Using Hand Gestures.
10.17148/IJIREEICE.2025.135018

[10] I. Yang, and S. Khan, (2022). Audio Player Controlling Based on Hand Gesture Technique.
10.1109/ICICTA54532.2022.9945876

[11] P. Sharma, and R. Tiwari, (2022). Gesture-Based Music Player Controller.
10.1109/ICITEE55344.2022.9971345

[12] K. Mehta, and R. Kapoor, (2024). Real-Time Emotion Based Music Player.
10.1016/j.procs.2024.05.112

[13] H. Gupta, and P. Das, (2025). Real-Time Facial Emotion Detection System Using Multi-Modal Fusion Deep Learning Architecture.
10.1109/ACCESS.2025.3467821

14. Gokila and Sivakumar, HYBRID DEEP LEARNING TECHNIQUES FOR ENHANCED ASPECT BASED SENTIMENT ANALYSIS AND CONTEXTUAL FEATURE OPTIMIZATION(2023), Volume 38 No. 11s,PP:1605-1629