RESEARCH ARTICLE OPEN ACCESS

FACE MOOD JUKE BOX

[1]Yerraguntla Syamala, [2]Mrs.CH. Radha Kumari
[1]Student, Dept. of Master of Computer Applications, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.
[2]Asst.Prof, Dept. of Computer Science & Engineering, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.

ABSTRACT

Music has a huge impact on our lives, it makes us feel incredible and can make every situation that little bit better. Lively music gets us going during a workout, and chill songs help us relax and unwind. Recognising the relation between music and mood, music recommendation systems have been proposed. But these systems tend to take into account elements such as listening history or genre not always indicative of a user's here-and-now emotional state.

Here, we present such a novel jukebox recommendation system that will make use of sophisticated facial emotion detection. The system provides personalised song recommendations according to a user's real-time emotions through the use of the AI-based facial emotion recognition tool, Gemini-provision. For instance, if someone is feeling happy, the system might interpret their facial expressions and propose an upbeat playlist to elevate their mood. Instead, if a user sounded depressed, the system would play soothing music.

The system is also capable of incorporating the user's language preferences, thus ensuring that recommendations are not only consistent with their emotion, rather that it is also consistent with their geographical origin and their musical preferences. This paper describes the system, explains the functionality of Gemini-provision, and discusses its architecture. We also discuss the advantages of this approach, in terms of the possibility of allowing the generation of a stronger relationship between music and an emotional status.

Keywords: *Gemini-pro-vision, Facial emotion detection, Listening Experience, Emotional state, Music recommendation.*

1. INTRODUCTION

This paper works on an individualized music recommendation system with facial emotion recognition. Music makes us feel emotions, and by reading facial expressions through a camera the system can detect emotions like happiness, sadness or anger and suggest music that fits how the user feels. Unlike the typical jukebox system that prompts the users at any time to select songs and play them in their preferred sequence or the pre-determined playlist playing system, the method provides a more customized and dynamic experience. The system, for instance, when detecting sadness, might recommend calming ballads; when detecting excitement, you will get energetic tunes. The platform is powered by Gemini-pro-vision, an advanced AI solution for precise emotion recognition that allows the music not only to be appropriate for the user's emotional mood, but also the language and culture of the consumer. This solution truly enriches the connection of the user with music by taking into account their emotional requirements at any particular time.

1.1 Emotion detection

Emotion detection is a subject of computer science that uses technology to understand how we feel. Such a system generally focuses on facial expressions although it may also integrate other features, such as voice characteristics and body movements. Picture an advanced machine learning algorithm that has been fed a tremendous amount of data of human faces of various emotional states. The system learns to recognize emotions like happiness (which includes large smiles and crinkled eyes), sadness (manifests itself in turned down mouth corners and furrowed brows), or anger (which includes a clenched jaw and narrowed eyes) by analyzing the minute muscle movements around the eyes, mouth and brow.

Although emotion detection is still improving, it is also not perfect. The video system can be adversely affected by cultural differences and personal facial mannerisms. However, in a jukebox project, emotion detection provides a means of understanding the user's feeling states and to match music recommendation based on a user's mood.

Emotion detection procedure mainly involves three prominent steps:

1. **Pre-processing:** It is the preparation of data, like croping, removing the effect of the light, transitioning the image into the quality where the system can interpret properly.
2. **Feature Extraction:** Once the pre-processing of the face is done, the system extracts features of the face relevant to emotions. This can apply to unique places such as the corners of the mouth, between the eyebrows, or the crinkling around the eyes. The algorithm then processes the position and motion of these features.
3. **Classification:** After the features are composed, the system classifies the emotion by contrasting the observed features with a large database of already labeled emotional expressions. Using sophisticated algorithms, it recognises the emotion that most closely corresponds to the analyzed facial features.

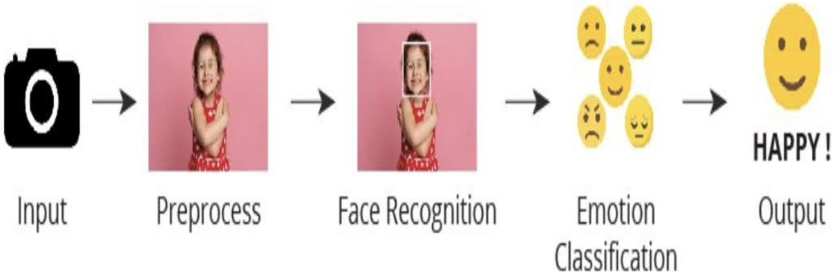


Fig 1:Proccedure

1.2 Language preference of music

Listeners' preferences for language in music can differ significantly. Some individuals prefer to listen to music in their native language, as it allows for a deeper connection with the lyrics and emotions expressed. Others may focus more on melody and rhythm, enjoying music regardless of the language. Additionally, some listeners may find music in foreign languages more intriguing or emotionally resonant. Ultimately, there is no right or wrong approach—it's all about personal taste. By offering options in Telugu and Hindi, this project allows users to select music based on their language preference, which could lead to a more meaningful connection with the recommended songs.

1.3 Music Recommendation

If the jukebox recognizes the man sitting in front of it is feeling sullen by reading his face, it can recommend similar mood songs taking into account his preferred language. So, if the jukebox senses a melancholic person who is a fan of Telugu music, then it can suggest some sad Telugu songs. If, on the other hand, the man or woman who loves nothing more than Hindi music is seen laughing, it’s safe to bet that the assistant is playing up-tempo Bollywood numbers. This customization customizes the music selection based on the listener’s emotions as well as the language preference contributing to the overall music experience.

1.4 Problem Statement

Conventional jukebox apparatuses cannot sense users' feelings therein and then play back music in response to the users' feelings, so known jukebox players has used fixed playlists or required users' instructions. In this project, we propose a relatively simple recommendation system for jukebox in which the user’s emotion is recognized by the Gemini-pro vision engine with the VideoDRiver’s image collection. The recommended songs will be based on the identified mood and will be in language that are preferred by the user (for example, songs in Telugu or Hindi), thus personalizing user's music experience. The objective is to develop an emotion-aware music recommendation system that enables a deeper connection with the user's music selection.

2. Literature Review

Mehendale, Ninad, et al.

Emotion recognition through facial expression, a natural human behavior for humans is yet not natural for machines. In this paper we propose a two-level CNN model for the problem of emotion recognition using facial expressions. [1]The facial feature vectors are extracted and removed background elements in the model to increase accuracy. Trained on a dataset that finds 154 subjects across 10,000 natural photos, the model IDs five facial expressions accurately 96% of the time.[1] Performance of the FERC system is improved by a novel background removal procedure, as well as its potential use in applications such as lie detection and predictive learning.[1]

Pandeya, Yagya Raj, et al.

This work addresses the problem of emotion in music video, which is a challenging task with fewer annotated data. The work demonstrates that when using CNNbased unimodal and multimodal scenarios, the combination of modalities does result in an increase in emotion classification accuracy by balancing dataset.[2] The most accurate model obtains an accuracy of 88.56%, showing promise for CNN-based multimodal models in emotion recognition over very few data.[3], [4]

Sajjad, Muhammad, et al.

This work focus of human behavior analysis via facial expression in the multimedia databases, which is useful in surveillance, medicine, sport, entertainment.[4], [5] The approach comprises face detection, tracking and recognition using a shallow CNN, as well as data augmentation to cope with illumination difficulties.[5] The method yields better performance in spontaneous facial expression recognition, as evidenced by subjective and objective testing.[5]

Kang, Yuhao, et al.

Through big data and affective computing, we proposed a method to analyse emotions from geo-referenced photos taken in tourist sites.[6] The system uses social media feeds and sophisticated computer vision to measure emotions expressed on the face. The study reveals the connection between the environmental context and emotions, and provides implications on geography-oriented emotion analysis.[6]

Li, Shan, Weihong Deng, et al.

In this paper, we have conducted a survey on the literature where deep learning techniques have been applied which includes expression unrelated challenges and overfitting on FER. [7], [8]This paper reviews the literature of data sets, algorithms and workflow of the standard deep FER system, and compares the advantages and disadvantages of state-of-the-art networks. It also discusses future directions to advance robust deep FER systems in live conditions.[9]

3. Scope and Objective

3.1 Scope

The main object of the project is to develop new jukebox recommendation system based on the use of facial expression based emotion detection combined with music scores. A camera interface is included within the system to film and analyze the users' facial expression concurrently. It will detect and classify a wide range of emotions (happiness, sadness, anger, etc..) on the fly. Using this emotion data, the system will suggest music tracks that are personalized to the user's current mood and taste, including options such as language (e.g., Telugu or Hindi). The project will include developing emotion detection and music recommendation algorithms, as well as producing an easy to use jukebox interface.

3.2 Objective

This project aims to revolutionize the original jukebox on vehicles and give users a new way of receiving the personalized music recommendations while they tell their own emotional state. *The key objectives are*:
A real-time detection and expression classification based fine-grained facial emotion detection algorithm.

- Develop a highly sophisticated music recommendation system, based on facial detection, that uses emotional data to recommend songs based on the user's mood.
- Build a Jukebox interface where these algorithm can be used in an intuitive and easy-to-use way, giving the user a fun and fluent experience.
- Facilitate inclusivity and accessibility by addressing varied users' emotional needs and language preferences in ways that augment and otherwise enrich their music-listening experience.

4. System Analysis

4.1 Existing System

The use of CNNs to solve face-Recognition problems has established itself as the approach for the recognition of human's emotional categories from the facial expressions. CNNs are a type of deep learning neural network and are excellent with visual data, which explains why they are used for image classification or object detection.

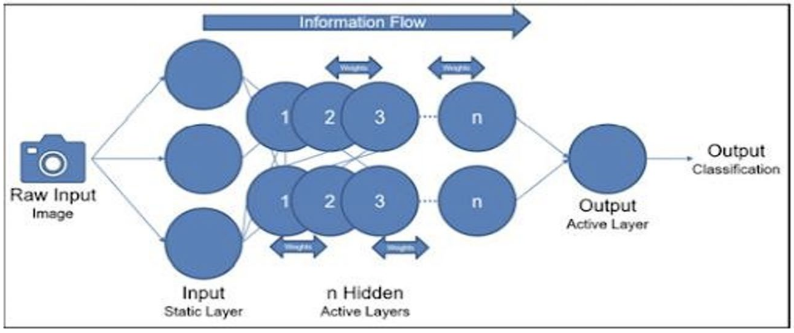


Fig 2:Exist-ing System

The standard procedure of FER through CNNs mainly consist of the subsequent steps:

1. **Collected Data and Preprocess**:
   We gather a large face dataset where each face image is labeled with one of multiple expression labels. During training, the images have been pre-processed to enhance their appearance, and to remove noise which could interfere with the learning from training images.
2. **Training the CNN Model**:
   The preprocessed face images are used to train the CNN model. At this stage, the CNN is optimized to pick up on-the-fly those input images, and the corresponding emotional labels (e.g., happiness, sadness, excitement, anger) anger). The training is based on the stacked convolution and subsampling layers structure of the network, which makes it possible to learn features hierarchically. the composition and the spatial relationship in the face data.
3. **Validation and Fine-Tuning**:
   The model is then evaluated on a held-out validation data set after training. This enables the system to measure its performance and to generalize well on new data. Depending on the evaluation, an altering of the model (fine tuning/optimization) may be possible to achieve better results.
4. **Real-Time Emotion Detection**:
   After training and validating, the model will be applied for real-time emotion detection. In this step, CNN is inputted with the images of human face from cameras or other imaging devices. Next, the model extracts facial features, such as eye gaze, eyebrow position, mouths shape, to predict the person's emotion.
5. **Output and Integration**:
   A system's output is a set of these predicted emotional labels for each input image. These estimates may be used in a variety of applications including recommendation systems, interactive user interface systems and surveillance systems. This conflation enables greater insights into the human, as well as a more personalized experience across a wide range of applications.

4.2 Proposed System

The model for the method for detecting emotion and suggesting music according to the present invention is described below:

1. **Input Capture:** The system starts with real-time input from a webcam to capture the facial expression of the user. These are the livestream data which are the primary inputs for emotion detection.
2. **Emotion Detection with the Gemini-pro Vision Model:** An exclusive AI engine from Amaryllo, the Gemini-pro Vision model, which accurately detects your emotions by analysing the facial image. It achieves this by utilizing the user's facial features and facial expressions using state-of-the-art deep learning algorithms to pin point the user's emotional state.
3. **Music Recommendation Engine Integration:** Now that we have the detected user and what language they would prefer, this' is sent to the music recommendation engine. Behind the system are complex algorithms that receive your mood and your choice of language in order to deliver "a playlist made especially for you."
4. **Integrate with Music Recommendation Engine**: As soon as the user's mood and the language he/she prefers are detected, they are transmitted to the Music recommendation engine. Their engine employs sophisticated algorithms to create a playlist catered to the user's mood and language of preference.
5. **a Music recommendation:** The music recommendation engine gives You a personalized playlist based on Your mood and also your language preference. For instance, if the current mood of the user is happy and the user likes the Hindi songs, the system may recommend lively Bollywood songs that indicate happy mood states.

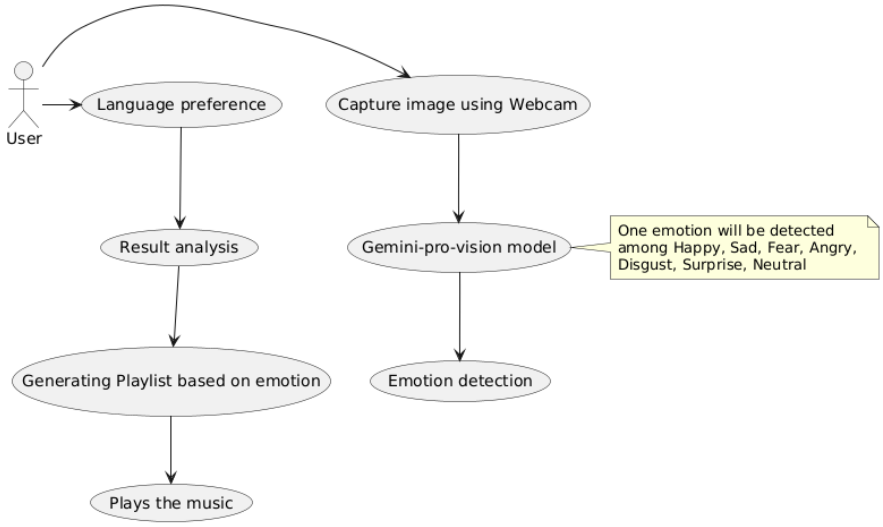

User
Language preference
Capture image using Webcam
Result analysis
Gemini-pro-vision model
One emotion will be detected among Happy, Sad, Fear, Angry, Disgust, Surprise, Neutral
Generating Playlist based on emotion
Emotion detection
Plays the music

Fig 3:Proposed System

4.3 Advantages of Proposed System

- **Efficiency**: As Gemini-Pro can be efficiently optimized for faster inference with fewer computational resources than conventional CNNs. This renders it convenient for real-time application, especially for mobile and edge computing systems.
- **Accuracy**: Gemini-Pro is more promising in achieving an improved performance in emotion identification tasks when compared to traditional CNN structures. This may be due to the utilization of more sophisticated methods as well as its handcraft with focus on emotion recognition.
- **Dedicated Attention to Emotion Detection**: In contrast to GP-based CNNs, which could demand more tweaking, Gemini-Pro is presumably tailored for the task of emotion detection. This specialized process may lead to stronger emotion recognition performance.
- **Scalability**: The construction of Gemini-Pro may be designed to scale easily to accommodate larger data or more demanding emotion recognition tasks.

4.4 Software Specifications

Component	Specification
Operating System	Windows 10, 11
Web Technologies	HTML, CSS, JavaScript
Programming Language	Python
Web Application Framework	FastAPI
Gemini API Key	Access to the Gemini-pro vision model for emotion detection

Table 1:Software Req

5. System Design

5.1 System Architecture

The system accepts the input image from user and which captured through the webcam and user's preference (Telugu or Hindi) is also accepted at the same time. The system then operates on these inputs to determine the current mood of the user through analysis with image recognition algorithms. According to the determined mood and the selected language, the system suggests recommended songs suitable for the user's emotion and language. This constitutes the fundamental working flow of the architecture of the system, described as follows:

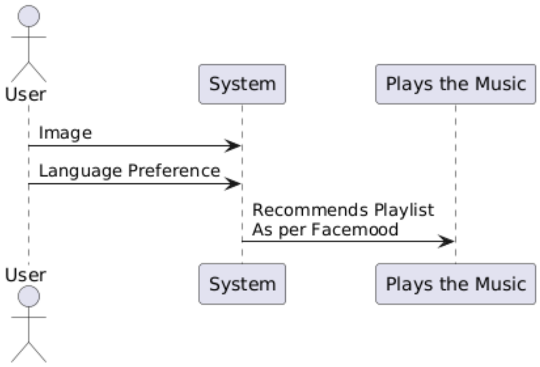


Fig 4: System Architecture

5.2 Module Description

1. **Input Module**:
   - **Webcam**: Takes the user's photo for facial emotion detection.
   - **Language Selection:** User is prompted to select music language preference, which will be saved as system preference.

2. **Facial Emotion Detection Model**:
   - Utilizes the "Gemini-Pro Vision" AI model to detect the user's emotion from the captured photo.
   - Constructed in Python via FastAPI with client authentication via a Gemini API key for model access.

3. **System Backend**:
   - Feeds the captured photo into the emotion detection model.
   - Gets the user language preference.
   - Uses natural language in combination with sentiment to make music suggestions.

4. **Recommendation Engine**:
   - Asks a music service to recommend a song by emotion and language.
   - Applies some algorithms (e.g., collaborative filtering, content-based filtering, or hybrid approaches) to make recommendations for users.

5. **User Interface**:
   - Shows the detected emotion & suggested songs in the user's native language.
   - Enables user interaction like Skipping tracks or replaying tracks and also user can change personal preferences.

6. **Integration with Jukebox**:
   - Embeds the recommendation engine into jukebox system for automatic song play.
   - Facilitates communication between the backend and jukebox for robust user experience.

5.3 UML Diagram

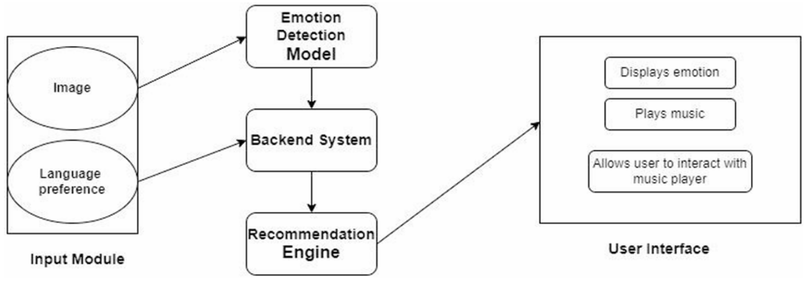


Fig 5: Use-Case::Diagram

6. PROPOSED METHODOLOGY

6.1 Basic Concept

A music player that can choose the songs according to how you are feeling! With Gemini-pro-vision, the system takes a look at your facial expression through a webcam or photo and tries to identify whether you're happy or sad. You then select your music language preference, and the system suggests songs that fit your mood and preferred language. For instance, if you tell it you're in a good mood and request English music, it may suggest peppy pop hits; if you're feeling mellow and want some Korean music, it might urge you to try soothing ballads.

6.2 Detecting Emotions with Gemini-Pro Vision

6.2.1 Google Generative AI

About Google Generative AI Google Generative AI is a set of machine learning models designed to assist in authoring text, playing music and composing images, which helps content generation, language translation, and reading comprehension by understanding large amounts of data to generate new content similar to what it has been trained on.[10], [11]

6.2.2 Large Language Model (LLM)

LLMs trained on massive datasets understand patterns in language, and can be used to generate text, translate languages, write creative content and answer questions.[12], [13] They are also applied in chatbots, machine translation, content generations and image captioning.[14], [15]

6.2.3 The Training Regime of a Large Language Vision Transformer

Training Gemini-pro-vision requires text-image pairs from large datasets which are fed and processed by powerful GPU's. The details of training data and training are proprietary, but those models are trained to correlate text and image information.[16], [17]

6.2.4 Gemini-Pro-Vision Model

Gemini-pro-vision is a generative AI model which can reason about both text and images, allowing it to be used for recognizing emotions from facial expressions.[18]

6.2.5 Google API Key And How to Create Gemini API Key

To work with Gemini-pro-vision, start by creating an API key in the Google Cloud Console, enabling the Vertex AI API, and then generating an integration key.[19], [20], [21]

6.2.6 Detecting Human Emotions with Gemini-Pro-Vision

Gemini-pro-vision's processing images of faces to determine peoples emotions including the facial feature eye gaze and mouth position.[22], [23]

6.2.7 System Prompt

A system prompt is an utterance used to inform a LLM what's the role and tone and style of the LLM, and even provide some background information. (It helps the AI respond, and also it makes the responses more accurate and creative by the AI.)[23]

7. IMPLEMENTATION

This section explains the implementation of a recommendation system of music insofar that it recommnds music according to the user's affective state by detecting their facial emotion and language-preference.

7.1 Model Used

The system is based on Geminipro Vision, a wide coverage large language vision model developed by Google AI, to predict user's dominant facial expressions (e.g. smiling, frowning, angry or sad etc) from images (taken by webcam or uploaded by the user).

7.2 Approach of the Model

We adopt the deep learning based facial emotion recognition technique, Gemini-pro Vision, trained on a large amount of labeled images. Model will extract facial features as well as circumstances that help in predicting the emotional status of the users accurately.

7.3 Phases of Development

1. **Data Acquisition and Pre-Processing:** Collect labelled images having expressions.
2. **Model Training:** Train Gemini-pro Vision on the dataset.
3. **API:** Create a communication mechanism between front-end and back-end using an API.
4. **Recommendation Engine**: Build a system that recommends music using an emotion's detection and language processing.
5. **Front-End**: Build a form for users to upload their own images, which is paired with entering their language preference.

7.4 The Google Generative AI Library

This project is an emotion detector using **Gemini-pro Vision**, and there is no element of music generation involved with Google's generative AI libraries.

7.5 System Integration to the Front-End

The procedure is as follows:

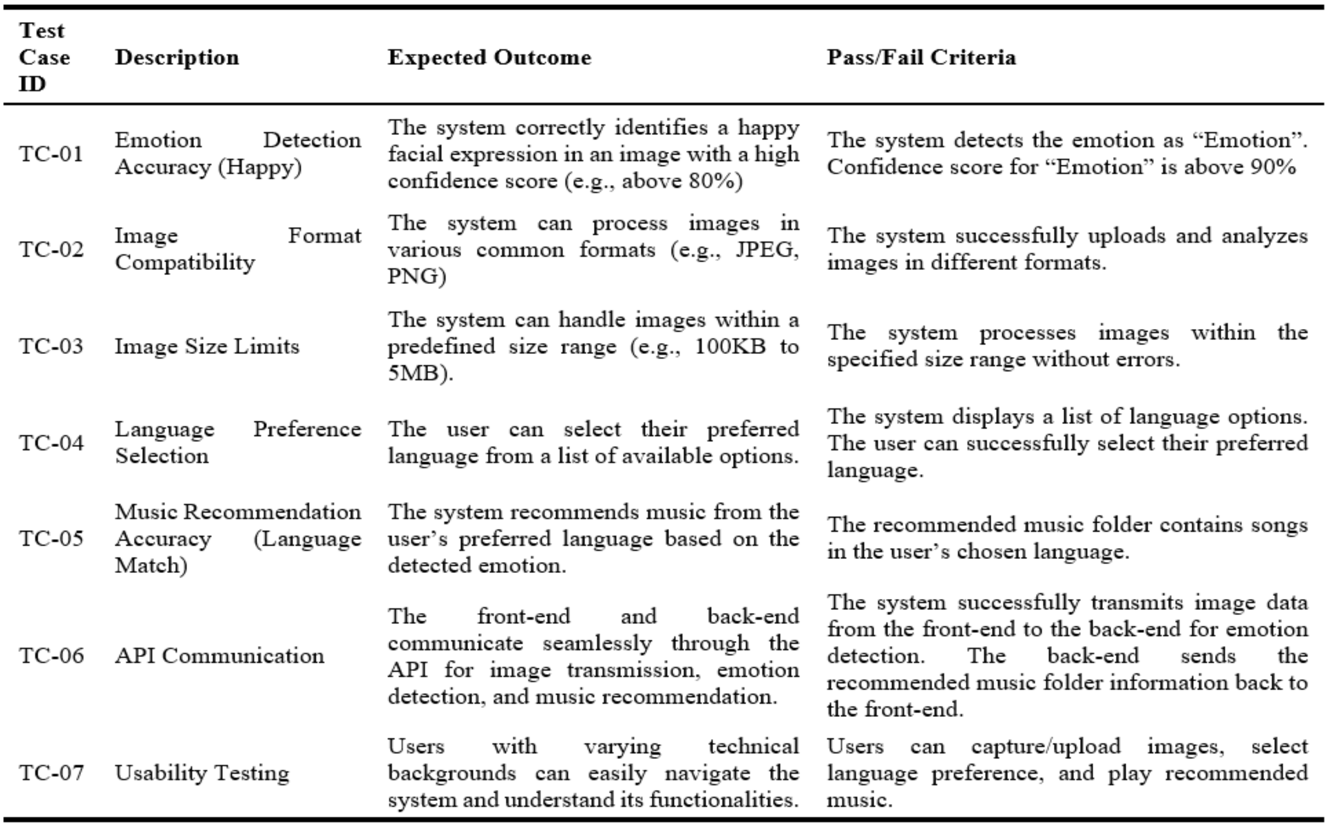
- Theimage is captured or otherwise recieved at the front, and is sent to the back via the API.
- The back-end recognizes the emotional state of the user and retrieves the language preference.
- On the front end, the recommendation engine chooses appropriate songs, which are returned to the front end to be played.

7.6 Testing and Feedback

Testing includes:

- **Emotion Recognition**: Validating the accuracy of the model with various facial expressions. Testing model accuracy with different facial expressions.
- Recommendation Engine To know the proper song to be fetched using emotion and language.
- Usability – User experience test.

7.7 Test Cases



Test Case ID	Description	Expected Outcome	Pass/Fail Criteria
TC-01	Emotion Detection Accuracy (Happy)	The system correctly identifies a happy facial expression in an image with a high confidence score (e.g., above 80%)	The system detects the emotion as "Emotion". Confidence score for "Emotion" is above 90%
TC-02	Image Format Compatibility	The system can process images in various common formats (e.g., JPEG, PNG)	The system successfully uploads and analyzes images in different formats.
TC-03	Image Size Limits	The system can handle images within a predefined size range (e.g., 100KB to 5MB).	The system processes images within the specified size range without errors.
TC-04	Language Preference Selection	The user can select their preferred language from a list of available options.	The system displays a list of language options. The user can successfully select their preferred language.
TC-05	Music Recommendation Accuracy (Language Match)	The system recommends music from the user's preferred language based on the detected emotion.	The recommended music folder contains songs in the user's chosen language.
TC-06	API Communication	The front-end and back-end communicate seamlessly through the API for image transmission, emotion detection, and music recommendation.	The system successfully transmits image data from the front-end to the back-end for emotion detection. The back-end sends the recommended music folder information back to the front-end.
TC-07	Usability Testing	Users with varying technical backgrounds can easily navigate the system and understand its functionalities.	Users can capture/upload images, select language preference, and play recommended music.

Table 2: Test Cases

8. CONCLUSION

We present a novel jukebox-based recommendation system that leverages facial emotion analysis to personalize music recommendation by integrating our Gemini-pro-vision towards deep emotional recognition when compared to popular CNN structures for facial emotion detection. This forms playlists based on users' emotions and the genre they prefer and as such this tool to me appears to be more personalized reflecting the mood and culture of the user. The feasibility of this approach will be evaluated through user assessments and comparison to other systems.

The potential uses of this emotion detection system are broad, much wider than music recommendations. In the realm of education, it could be revolutionary by measuring student interest and adapting course material. In customer service, it might usher in more empathetic interactions by reading customer emotions. It also could serve as an early warning system for mental health issues, improve the gaming experience by altering gameplay based on emotions and help more precise targeting of advertisements by gauging audience reaction.

"This research is a big step in the direction of more sophisticated, expressive algorithms to help us search, discover, and explore not only music, but also information, images, and videos," she added, "Because all of these, you can find with emotion."

ACKNOWLEDGMENT

I am thankful to the Management of Amrita Sai Institute of Science and Technology for giving me an opportunity to work with his project.

I would like to the thank **Dr. M. Sasidhar**, Principal, Amrita Sai institute of science and technology, for his constant encouragement and support during the progress of this work.

I am deeply grateful to **Dr. P. Chiranjeevi**, Professor and Head of the Department, for his valuable guidance and consistent support during the course of the project.

A special note of thanks to my internal guide, **Mrs.CH. Radha Kumari**, for her exceptional guidance, constant motivation, and continuous encouragement, which played a crucial role in the successful completion of this project.

YERRAGUNTLA SYAMALA

REFERENCES

[1] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," SN Applied Sciences, vol. 2, no. 3, Feb. 2020, doi: 10.1007/s42452-020-2234-1.

[2] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Music video emotion classification using slow–fast audio–video network and unsupervised feature representation," Scientific Reports, vol. 11, no. 1, Oct. 2021, doi: 10.1038/s41598-021-98856-2.

[3] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Deep-Learning-Based Multimodal Emotion Classification for Music Videos," Sensors, vol. 21, no. 14, p. 4927, Jul. 2021, doi: 10.3390/s21144927.

[4] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimedia Tools and Applications, vol. 80, no. 2, p. 2887, Sep. 2020, doi: 10.1007/s11042-020-08836-3.

[5] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, "Human Behavior Understanding in Big Multimedia Data Using CNN based Facial Expression Recognition," Mobile Networks and Applications, vol. 25, no. 4, p. 1611, Sep. 2019, doi: 10.1007/s11036-019-01366-9.

[6] Y. Kang et al., "Extracting human emotions at different places based on facial expressions and spatial clustering analysis," Transactions in GIS, vol. 23, no. 3, p. 450, Jun. 2019, doi: 10.1111/tgis.12552.

[7] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial Expression Recognition using Visual Saliency and Deep Learning," arXiv (Cornell University), Jan. 2017, doi: 10.48550/arXiv.1708.08016.

[8] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," Procedia Computer Science, vol. 175, p. 689, Jan. 2020, doi: 10.1016/j.procs.2020.07.101.

[9] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, p. 1195, Mar. 2020, doi: 10.1109/taffc.2020.2981446.

[10] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," Artificial Intelligence Review, vol. 55, no. 1. Springer Science+Business Media, p. 589, Jul. 02, 2021. doi: 10.1007/s10462-021-10039-7.

[11] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Šeďová, "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arXiv.2301.04246.

[12] S. Chakraborty, "Generative AI in Modern Education Society," arXiv (Cornell University), Dec. 2024, doi: 10.48550/arxiv.2412.08666.

[13] Q. Wang et al., "Attention Paper: How Generative AI Reshapes Digital Shadow Industry?," p. 143, Jul. 2023, doi: 10.1145/3603165.3607442.

[14] H. Gilbert, M. Sandborn, D. C. Schmidt, J. Spencer-Smith, and J. White, "Semantic Compression with Large Language Models," Nov. 2023, doi: 10.1109/snams60348.2023.10375400.

[15] H. Naveed et al., "A Comprehensive Overview of Large Language Models," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arXiv.2307.06435.

[16] G. Team et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," arXiv (Cornell University), Mar. 2024, doi: 10.48550/arXiv.2403.05530.

[17] Y. Shen et al., "MoViT: Memorizing Vision Transformers for Medical Image Analysis," Lecture notes in computer science, p. 205, Oct. 2023, doi: 10.1007/978-3-031-45676-3_21.

[18] "Gemini API reference." May 2025. Accessed: May 15, 2025. [Online]. Available: https://ai.google.dev/gemini-api/docs/api-overview

[19] N. Kumari, "Using Gemini in BigQuery for sentiment analysis." Jun. 2024. Accessed: May 15, 2025. [Online]. Available: https://cloud.google.com/blog/products/data-analytics/using-gemini-in-bigquery-for-sentiment-analysis

[20] "A Comprehensive Guide to Access Google's New Gemini AI API with AIsBreaker." Jan. 2024. Accessed: May 15, 2025. [Online]. Available: https://aisbreaker.org/blog/2024-01-13-use-google-vertexai-gemini

[21] "Image understanding." Apr. 2025. Accessed: May 15, 2025. [Online]. Available: https://ai.google.dev/gemini-api/docs/vision

[22] "Google AI Studio vs. Vertex AI vs. Gemini." May 2025. Accessed: May 15, 2025. [Online]. Available: https://cloud.google.com/ai/gemini

[23] "Gemini API." Apr. 2025. Accessed: May 15, 2025. [Online]. Available: https://ai.google.dev/gemini-api/docs