International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Facial Emotion-Based Song Recommender System Using CNN

Jauwad Jamal, Prince Yadav
SCSE Galgotias University, Greater Noida
jauwadjamal5@gmail.com, princey292.py@gmail.com

ABSTRACT

Music has an impact on human brain activity. Users can sustain a specific emotional state with the aid of an emotion-based music player with an automatic playlist. An emotion-based music player that builds playlists from user images is suggested by this study.

This process has been automated using a variety of algorithms. Nevertheless, current algorithms are slow, need more hardware to operate, and are incredibly inaccurate. This paper presents an algorithm that not only automates the process of generating an audio playlist, but also to classify those songs which are newly added and the main task is tocapture current mood of person and to play song accordingly. This enhances the system's efficiency, faster and automatic. The main goal is to reduce the overall computational time and the cost of the designed system. It also aims at increasing the accuracy of the system. The most important goal is to make change the mood of person if it isa negative one such as sad, depressed. This model is validated by testing the system against user dependent and user independent dataset.

KEYWORDS: Convolution neural network, Long Short-term memory, Emotion detection, audio classification, hidden layers, Max-pooling.

INTRODUCTION

Human emotion recognition and expression are critical components of the communication system. Emotions are something that humans can express and identify. Computers use sensors or image analysis to try and determine human emotions. We communicate with a lot of people in our daily lives and at work, either directly or indirectly over the phone. In certain situations, it is important for people to be conscious of the emotions of the person they are speaking with. The following categories apply to human emotions: surprise, fear, anger, happiness, sadness, disgust, and neutral.
With the exponential growth in digital content, recommender systems have become an integral part of user experience across platforms. Emotion-aware recommendation systems represent a paradigm shift in personalization, where the emotional state of the user is considered for content delivery. This paper explores the integration of facial emotion recognition with music recommendation, enhancing user engagement by aligning song suggestions with the user's current mood. Emotion recognit...
Emotions are largely communicated through speech and facial expression. The energy in a speaking utterance is conveyed by the body language and facial tone, which can be initially altered to convey various emotions. Together with the information received by any other sensory organs, humans are able to discern these signal variations with ease. This research examines how emotions can be captured by words, images, or sensors.

Since it is a significant kind of enjoyment for listeners and music lovers and occasionally even offers a therapeutic approach, music plays a crucial role in improving a person's quality of life. Music may simultaneously and gradually transform a person's bad emotions into positive ones because it communicates where words cannot.

Body language, voice, facial expressions, gestures, and more can all be used to convey emotions. We employ facial expressions to help the system interpret the user's mood. We can record the user's face expression with the camera on the mobile device. There are numerous

LITERATURE REVIEW

Emotions can be expressed via body language, voice, gestures, facial expressions, and more. To assist the algorithm in determining the user's mood, we use facial expressions. The mobile device's camera allows us to capture the user's facial expression. A variety of emotion identification technologies are available to detect an emotion in a captured image. For this application, neural networks are being used to identify emotions.

The user wouldn't have to waste time looking for tunes or conducting searches. Three modules—the emotion extraction module, the audio extraction module, and the emotion-audio extraction module—were included in the suggested architecture. Despite many drawbacks, such as the suggested system's inability to accurately capture every emotion, because the images in the image dataset being used are less readily available. For the classifier to produce reliable results, the image that is given into it needs to be taken in a well-lit environment. In order for the classifier to accurately anticipate the user's sentiment, the image quality must be at least 320p. In the wild, handcrafted features frequently don't generalize well enough.

"Emotion Based Music Recommendation" was proposed by H. Immanuel James, J. James Anto Arnold, J. Maria Masilla Ruban, and M. Tamilarasan (2019). Its goal is to scan and understand facial emotions in order to create a playlist that reflects those emotions. Creating a suitable playlist based on a person's emotional characteristics eliminates the laborious process of manually separating or organizing songs into various categories. The goal of the suggested method is to create emotion-based music players by identifying human emotions. Face detection is done with a linear classifier. A facial landmark map of a given face image is produced using regression trees trained using a gradient boosting technique, based on the

emotion recognition systems that can identify an emotion from an image that has been taken. Neural networks are being used for this application in order to recognize emotions.

intensity values indexed of each pixel. Emotions are categorized using a

RESULT OF LITERATURE REVIEW

One of the most significant and crucial aspects of the project is the report's section on classifying emotions based on facial expressions. As a result, research papers, whitepapers, and earlier findings are published close to my study.

Extensive literature indicates a growing interest in affective computing and its application in entertainment technologies. Studies by Picard et al. (1997) laid foundational theories of emotion-enabled computing. Recent works have shown promising results using hybrid models combining CNNs and Recurrent Neural Networks (RNNs), especially LSTM architectures, for dynamic emotion recognition from video feeds.

Empirical evaluation of various classification algorithms indicates that CNNs outperform traditional SVMs and Decision Trees in accuracy and generalization. The importance of well-annotated datasets and lighting normalization has also been emphasized, as real-time applications suffer from variations in environmental conditions.

based on some of the department's scientists. Given that researchers attempted to create a new app with a far more effective. They began by grouping emotions into four main categories—happiness, sadness, anger, neutral, fear, surprise, and disgust—using a computationally efficient program. Other emotions, such as grief, are a combination of sadness and anger. Because convolutional neural networks (CNNs) may provide good accuracy and precision in a reasonable period of time, this proposed project can use CNNs for face expression identification.

multiclass SVM classifier. There are four categories of emotions: surprise, sadness, anger, and happiness.

HARDWARE AND TOOLS USED

a) Wireshark Designing Tool: Figma
c) yEd Live: State Transition Designing Tool
c) yEd Live, Breakdown's Designing Tool
d) OpenCV is the suggested image capture tool.
e) Convolutional Neural Network is the suggested algorithmic tool (for emotion detection).
f) Python (Server Layer) and Flutter (Mobile Application Layer) are the preferred programming languages.
g) Favorite Platform: iOS and Android smartphones.

IMPLEMENTATION

The suggested algorithm centers on an automatic music recommendation system that plays music based on the user's current emotional state or mood. Every time the application opens, a picture of the user is taken; as a result, the user's current emotion is recorded and identified. The graphic provides information that suggests the song being performed in relation to the feeling. The system comprises multiple integrated modules, each designed to handle specific subtasks in the pipeline. The facial expression recognition component utilizes Haar Cascades for face detection and a pre-trained CNN model for emotion classification. The CNN is trained on FER-2013 and CK+ datasets to identify emotions like happiness, sadness, anger, fear, surprise, disgust, and neutral. For music recommendation, songs are tagged based on tempo, key, and lyrical content.

All of the songs on the phone are already divided into seven categories: neutral, fear, disgust, anger, surprise, happy, sad, and sad. Additionally, the recently uploaded music are categorized dynamically to suit the mood. It is made up of three modules: one for system integration, one for facial expression identification, and one for song Audio Extraction Module: Following the extraction of the user's emotion, the user is presented with music or audio that reflects the emotion they spoke. A list of songs that are based on the emotion is then presented, and the user is free to listen to any song they choose. The songs are emotion recognition. The modules for auditory emotion recognition and facial expression recognition are mutually exclusive.

A. Data Collection

One by one, the Raw dataset for each of the seven emotions is retrieved from Google Images. To recognize facial expressions, additional data is extracted from Kaggle datasets.

B. Dataset with Training

The training and testing stages are completed before the model is processed. Datasets that are taught to the model or that learn are known as trained datasets.

The system uses a dataset of faces (pictures) with their corresponding expressions during training; the eye should be primarily in the center. It then learns a set of weights that divide the facial expressions for categorization.

Emotion Extraction Module: A camera or webcam is used to take a picture of the user. To enhance the performance of the classifier, which is used to identify the face in the image, the frame of the webcam feed is transformed to a grayscale image after it has been captured. After the conversion is finished, The picture is fed into the classifier system, which uses feature extraction methods to identify the face in the webcam feed's frame. To identify the user's expressed mood, the trained network receives individual attributes from the extracted face.

These pictures will be used to train the classifier so that, using the information it has already learned from the training set, it can determine the location of facial landmarks in a brand-new, unknown collection of pictures when it is shown them. the new facial landmarks' coordinates that it identified. The large data set from CK is used to train the network. This is used to determine the emotion that the user is expressing.

shown in that order based on how frequently
The user would listen to them. Web technologies including PHP, MySQL, HTML, CSS, and JavaScript were used in the development of this module. The Emotion-Audio Integration Module stores the feelings that are derived from the songs, and the web page created using PHP and

Test Data

At the time of testing, classifier takes images of face with respective eye center locations, and it gives output as predicted expression by using the weights learned during training.
For recognizing an unknown image (testing), the sequence is:
1. Spatial normalization
2. Image cropping
3. Down-sampling
4. Intensity normalization
Figure 4: Architecture of proposed Convolution Neutral Network. It has five layers: first layer is convolution, second
layer is sub-sampling, the third layer is convolution, fourth layer is sub-sampling, fifth layer is fully connected layer
and final responsible for classifying facial image.

FINAL RESULTANT MODULE

To make preprocessing and detection easier, quicker, and more efficient, all of the images in the collection are first transformed to grayscale. Pixels make up each input image (e.g. 48x48). The convolution layers (hidden layers) now receive the images that are represented by pixels. Maximum pooling between each layer is carried out, with the intention of

This is done by down sampling the input data or image, which lowers the dimensions and makes it possible to infer features found in subregions. The purpose of this is to prevent over-fitting. Additionally, it lowers computing costs by learning fewer parameters. For instance, pooling is done between all hidden layers if the input picture has a matrix 4x4 representation and the desired output is 2x2. To avoid over-fitting, data is then routed to the dense layer. In neural networks, the dropout approach is used to lessen over-fitting. In the output layer, the identified

class. The next step is to choose any training dataset for the music model, assuming that the detected expression is cheerful. The dataset is currently being trained based on the music playing match. Songs are categorized using an LSTM neural network. To improve and speed up classification, one hot encoding is used to convert categorical variables into binary vectors.

CONCLUSION

The finest music player experience for the user is provided by the Emotion Based Music Player, which automates the process. The software meets all of the fundamental demands of music lovers without bothering them like other apps do. It makes use of technology in a variety of ways to improve system-user interaction. By taking the picture, it makes the user's job easier.
identifying their emotions through the phone's camera and recommending a personalized playlist with cutting-edge features. The song's transition from a low to an energetic tone gradually transforms the user's negative or poor ideas into happy ones.

FUTURE SCOPE

Google Play Music can be added to Music Player in the future, enabling speech-based access to the entire application and the ability to play songs that are not stored locally. Users who are searching for music based on their emotional behavior and mood would greatly benefit from the Emotion Based Music System. It will lessen the need for searching.

Future enhancements to this system may include integration with wearable devices to capture physiological signals (heart rate, galvanic skin response) for multimodal emotion detection. Furthermore, generative AI models can be incorporated to dynamically generate or modify music tracks based on detected emotional tone, offering a deeply immersive experience.

Another promising direction involves implementing reinforcement learning to allow the system to adapt its recommendations based on long-term user feedback. This will help not only in personalizing content but also in making the system more robust to user-specific emotional nuances.

time for music, cutting down on wasteful time and improving the system's overall precision and effectiveness. In addition to lowering physical stress, the technology will benefit music therapy programs and could help the patient receive treatment from a music therapist.

REFERENCES

[1] H. Immanuel James, J. James Anto Arnold, J. Maria Masilla Ruban, M. Tamilarasan, "Emotion Based Music Recommendation", 2019.

[2] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning”, Nature, 521(7553), 436–444, 2015.

[3] P. Ekman, “Basic emotions”, Handbook of Cognition and Emotion, 45–60, 1999.