Music recommendation system based on face emotion recognition using Convolution Recurrent Neural Network(CRNN)

A.Annie Micheal[1], Avanthika Suresh[2], Archita Prabhakar[3] and Dhanika Dennis[4]

[1,2,3,4]*Sathyabama Institute of Science and technology, Chennai, India*

Article Information

Received: 15-11-2023
Revised: 30-11-2023
Published: 15-12-2023

Keywords
Music recommendation system; Convolution Recurrent Neural Network; Face emotion recognition; Similarity distance

****Correspondence Email:***
annymick@gmail.com

Abstract

The emergence of commercial music streaming services, which can be accessed via mobile devices, has greatly enhanced the accessibility of digital music compared to previous periods. Organizing digital music can be a laborious and even overwhelming task, resulting in information overload. Therefore, the creation of an automated music recommender system that can effectively search music collections and offer suitable song recommendations to consumers is quite beneficial. By employing a music recommender system, music providers can predict and deliver appropriate song recommendations to users based on the audible characteristics of previously listened-to music. The aim of our study is to create and execute a music recommender system that leverages audio signal properties to offer recommendations based on similarity. This study leverages a convolutional recurrent neural network (CRNN) to extract features and employs a similarity distance metric to evaluate the similarity of these derived features. The study's findings indicate that consumers have a preference for recommendations that consider music genres rather than those that solely concentrate on similarities. The system's performance was assessed using the FER2023 dataset and compared to the most advanced techniques, resulting in an accuracy of 85%.

1. Introduction

A recommendation system operates as a filtering mechanism, assessing the perceived importance that a user would assign to a particular piece, such as a music. Large engines rely on recommender algorithms to provide projections and offer users a set of options. A music recommendation system can be classified into two main categories. The effectiveness of the content-based strategy relies on the degree of similarity between different items. Users generate playlists by selecting songs they like or dislike while using a streaming music service. The fundamental idea behind a content-based recommendation system is to analyses the description of the user's preferred music, identify its keywords, and subsequently compare them with the keywords of other songs. Based on this comparison, the system can offer new songs to the user. Our recommendation system utilizes collaborative filtering, which is based on the similarity of people' preferences. We employ matrices with ratings for each song to make these recommendations. The collaborative system is built around the convergence of users' shared preferences and music ratings. If consumers A and B have similar preferences, it is presumed that they can be recommended similar songs. Put simply, if user A finds a music enjoyable, it is probable that user B will also find it enjoyable, and vice versa. Collaborative recommendation systems are often considered to be more accurate because they depend on direct user interactions rather than content similarities.

2. Literature Review

[1] The study centersaround a system that recommends songs to users, taking into account their emotional state. This system use machine vision component to ascertain the user's emotion by analyzing facial expressions and interactions with the Chabot.Upon identifying the emotion, the system promptly recommends a song that corresponds to that emotion, thereby significantly reducing the time required for the user to manually pick and play music. The model is equipped with modules that can recognize facially expressed emotions and attitudes conveyed during a Chabot encounter, which enhances the effectiveness of the music recommender system. The model is equipped with modules that can recognize facially expressed emotions and attitudes conveyed during a Chabot encounter. These modules enhance the effectiveness of the music recommender system, making it more reliable and powerful.

[2] The primary objective of the study is to identify human emotions in order to create a music player that is based on emotion. The Emotion-based music program aims to assist anyone seeking music that is influenced by emotions and emotional conduct. It has the potential to decrease the duration of music searches, resulting in a reduction of superfluous processing time and ultimately enhancing the overall accuracy and efficiency of the system. The program addresses the fundamental requirements of music enthusiasts, while avoiding the inconveniences commonly seen in existing applications. It leverages technology to enhance the system's connection with the user through various means. The technology simplifies the task for the user by utilizing a camera to capture an image, analyzing the user's emotion, and recommending a personalized playlist from a Spotify Premium account through a sophisticated and interactive interface.

[3] The primary goal of this music recommendation system is to offer tailored suggestions to consumers based on their individual likes. Examining the face expression or user mood might provide insights on the user's present emotional or mental condition. Music and videos offer a great opportunity to provide a wide range of options to customers based on their preferences and past data. Humans commonly utilize facial expressions to enhance the clarity of their verbal communication and convey the intended meaning within a certain setting. Over 60 percent of users experience difficulty in selecting a certain song from their extensive music library due to its size. By creating a recommendation system, users can receive guidance on selecting music that can help them alleviate tension. The user can save time by effortlessly discovering songs that match their mood, since the system automatically detects and presents the most suitable tracks. A webcam is used to capture the user's image. A photo of the user is captured, and based on the user's mood or emotion, a suitable song from the user's playlist is displayed that aligns with the user's preference. The system was developed via the facial landmarks scheme and underwent testing across multiple scenarios to assess the achieved outcome. Observations indicate that the classifier achieves an accuracy exceeding 80 percent for the majority of test cases, which is considered a commendable level of accuracy for emotion classification. Furthermore, the classifier demonstrates the ability to precisely forecast the user's expression in a real-time situation during live testing.

[4] A prototype for a dynamic music recommendation system, which utilizes human emotions as a basis, has been created. The tunes for each emotion are trained based on individual human listening patterns. The combination of feature extraction and machine learning methods allows for the detection of emotions from real faces. Once the mood is determined from the input image, corresponding music will be played to engage the users. This strategy establishes a connection between the application and human emotions, hence providing a personalized experience for the users. Thus, our proposed system focuses on detecting human emotions in

order to construct an emotion-based music player employing computer vision and machine learning methodologies. We employ OpenCV for emotion identification and music recommendation in our experimental findings.

[5] A proposed recommender system aims to accurately detect a user's emotions and provide a curated collection of appropriate tunes that could potentially improve their mood. In order to provide consumers with a curated selection of music tracks that effectively uplift their spirits, we conducted a brief investigation to understand the immediate influence of music on mood. The proposed system employs emotion recognition to identify negative emotional states in individuals. In response, it automatically generates and plays a curated playlist of uplifting music to improve their mood. Alternatively, if a positive mood is detected, a suitable playlist will be provided, comprising various music genres that will enhance the enjoyable emotions. The recommender system is implemented using Principal Component Analysis (PCA) methodologies and the Fisher Face algorithm.

3. Proposed Methodology

The objective of a music recommendation system using emotion detection is to deliver individualized music suggestions that are specifically customized to the user's emotional state or preferences. Figure 1 shows the architecture diagram for the proposed methodology. This particular system utilizes emotion recognition technology to examine the emotional elements present in music and determine the emotional state or mood of the user. This determination is made by using a range of cues, including user inputs, biometric data, and contextual information.

In order to categorize the music genres, we employed the Convolutional Recurrent Neural Network (CRNN) [6] architectures. CRNNs are employed due to their ability to extract salient information for accurate prediction outcomes. In addition to examining frequency characteristics on the spectrogram, CRNNs are capable of analyzing temporal sequence patterns. Ultimately, the feature vectors generated prior to the classification layer can serve as a foundation for making recommendations. The CRNN design employed a dual-layered Recurrent Neural Network (RNN) [7] with Gated Recurrent Units (GRU)[8] to condense 2D temporal patterns derived from the outputs of four Convolutional Neural Network (CNN) layers [9]. The essential premise of this model is that Recurrent Neural Networks (RNNs) are more effective than Convolutional Neural Networks (CNNs) in aggregating temporal patterns. However, CNNs are still utilized on the input side for extracting local features. CRNN use RNNs to consolidate the temporal patterns, rather than employing methods such as averaging the outcomes from shorter segments or using convolution and sub-sampling as in other CNNs.
The CRNNs structure consists of the following components:

· Four layers of convolutional with kernel 3 × 3, feature maps (68-137-137-137), stride 1 and using padding to maintain input dimensions.

· Each convolutional layer employs batch normalization and ReLUactivation[10].

· Max-pooling layer for each convolutional layer with kernel ((2 × 2)-(3 × 3)-(4 × 4)-(4 × 4)) and same stride.

· Every convolutional layer incorporates dropout with a rate of 0.1.

· The model consists of two layers of Gated Recurrent Units (GRU) with 68 feature maps. The initial layer receives input data in the form of sequences and produces output data in the form of sequences. Within the second layer, the input data is shown as a sequential arrangement, whereas the outcome is represented as a singular value.

· The output layer utilizes the sigmoid function.

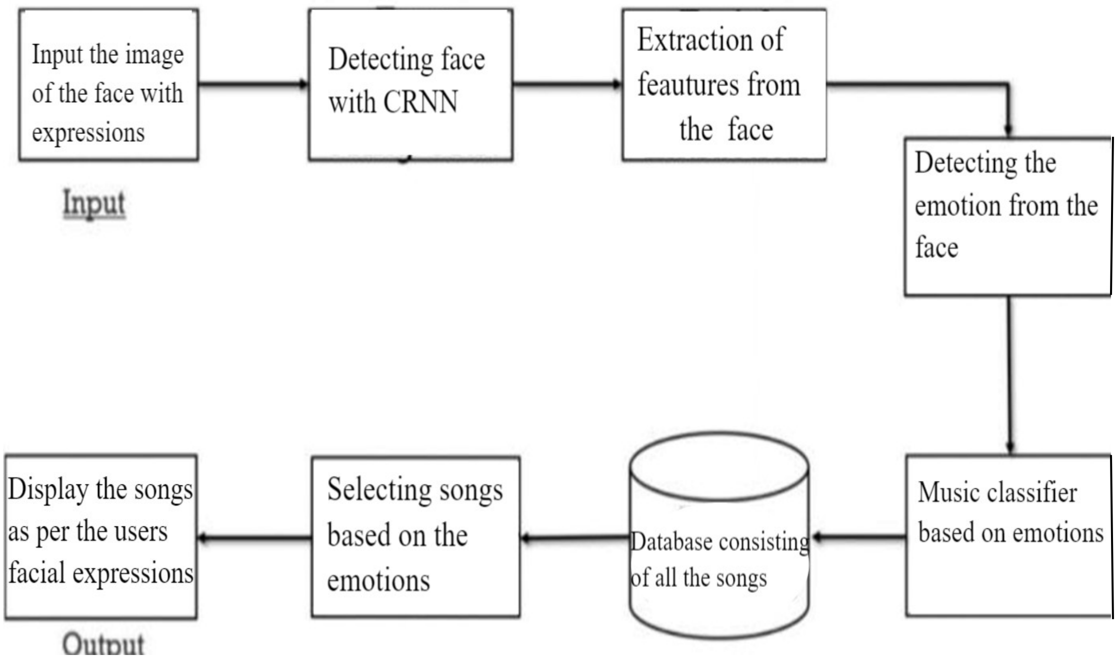
Input the image of the face with expressions
Input
Detecting face with CRNN
Extraction of feautures from the face
Detecting the emotion from the face
Music classifier based on emotions
Database consisting of all the songs
Selecting songs based on the emotions
Display the songs as per the users facial expressions
Output

Fig.1 Architecture diagram

4.Result and Discussion

For emotion detection, FER2013 dataset is used. This dataset consists of 38,887 48x48 sized grey scale face images with 7 emotion categories shown as in Table 1. The facial expression is divided into seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral) based on the emotion conveyed in the facial expression.

Table 1. Emotion labels in the dataset

Label	Emotion	No.of images
0	Angry	4593
1	Disgust	547
2	Fear	5121
3	Happy	8989
4	Sad	6077
5	Surprise	4002
6	Neutral	6198

Among the total of 38,887 photos, 28,709 were utilized for training the model, while the remaining images were allocated for testing purposes. A webcam is used to capture the facial image of the user. Each image undergoes preprocessing and resizing to a 48x48 array of grayscale values. It is then evaluated using the testing data. The image is preprocessed to identify the face using the Haar-based cascaded classifier within the OpenCV framework. The figure 2 displays some photographs from the FER2023 dataset.Table 2 presents a comprehensive assessment of the state-of-the-art methods for the FER2013 dataset. Figure 3 displays the result of the suggested system.

Fig. 2 Sample images from FER2023 dataset

Table 2. Comparison of the State-of-the-art methods

Paper	Method	Accuracy
Chidambaram, G., et al. [2]	Deep neural Network	69.14%
Yu, Z., et al[11]	convolutional neural network	62.5%
Athavle, Madhuri, et al. [3]	convolutional neural network	71%
Proposed system	CRNN	85%

5. Conclusion

To summarize, the incorporation of emotion recognition into a music recommendation system signifies a notable progress in the realm of artificial intelligence and data analysis. The primary objective of this innovative application is to improve users' enjoyment of music by providing tailored recommendations that correspond to their emotional preferences. This technology possesses the capacity to enhance the connection between individuals and the music they enjoy by understanding and effectively responding to the user's emotional state. This work utilizes convolutional recurrent neural networks (CRNNs) to extract features, then assesses the similarity between these obtained features using a similarity distance metric. The system's performance was evaluated using the FER2023 dataset and compared to state-of-the-art approaches, achieving an accuracy of 85%.

References

[1]Krupa, K. S., et al. "Emotion aware smart music recommender system using two level CNN." 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2020.

[2] Chidambaram, G., et al. "Music recommendation system using emotion recognition." International Research Journal of Engineering and Technology (IRJET) 8.07 (2021): 2395-0056..

[3] Athavle, Madhuri, et al. "Music recommendation based on face emotion recognition." Journal of Informatics Electrical and Electronics Engineering (JIEEE) 2.2 (2021): 1-11.

[4] Florence, S. Metilda, and M. Uma. "Emotional detection and music recommendation system based on user facial expression." IOP conference series: Materials science and engineering. Vol. 912.No. 6.IOP Publishing, 2020.

[5]Shalini, Shantha K., et al. "Facial Emotion Based Music Recommendation System using computer vision and machine learning techiniques." Turkish journal of computer and mathematics education 12.2 (2021): 912-917.

[6] Chatterjee, Chandra Churh. "An Approach Towards Convolutional Recurrent Neural Networks." Medium.url: https://towardsdatascience. com/an-approach-t owards-convolutional-recurrent-neural-networks-a2e6ce722b19 (visited on 12/28/2019) (2019).

[7] Schmidt, Robin M. "Recurrent neural networks (rnns): A gentle introduction and overview." arXiv preprint arXiv:1912.05911 (2019).

[8] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[9] O'Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[10] Agarap, Abien Fred. "Deep learning using rectified linear units (relu)." arXiv preprint arXiv:1803.08375 (2018).

[11] Yu, Z., Zhao, M., Wu, Y., Liu, P., & Chen, H. (2020, July). Research on automatic music recommendation algorithm based on facial micro-expression recognition. In 2020 39th Chinese control conference (CCC) (pp. 7257-7263). IEEE.