Emotion-Aware Music Recommendation System

Sai Suman Chitturi, Praneeth Kapila

Vasavi College of Engineering, Hyderabad, India.

Abstract: Music is an essential part of our regular life. It cheers us up and makes us feel better. Not all forms of music are appropriate for every mood. Furthermore, ever-growing digital music catalogues make it virtually impossible to recollect a specific tune that fits the present emotion. Besides that, due to the enormous number of songs accessible, people are frequently perplexed when selecting a track. This necessitates the development of a context-sensitive music recommendation system. Therefore, a context-aware music recommendation system is presented that assists in identifying the user's current emotion and suggesting music which is relevant to that emotion. A comprehensive strategy is presented to improve user preference prediction; the technique integrates context and emotion elements and strives to give users a more convenient, intuitive, and pleasurable listening experience. Finally, the evaluation and performance metrics and the results will be research.

Keywords: Convolutional Neural Networks, Deep Neural Networks, Multi-task Cascaded Neural Network, Clustering.

1. Introduction

Music is universal and easily accessible in our everyday lives, thanks to the tremendous growth of digital music technologies. People listen to music on various sources to improve their mood and enliven the environment. However, due to the enormous number of songs accessible, people are frequently perplexed when selecting a track.

People enjoy listening to different genres of music according to their mood. So, Naive recommendations do not work as intended. This necessitates the development of a context-sensitive music recommendation system. Context-aware recommendation systems also need accumulation of additional data, which can be difficult to come by but has shown to be worthwhile in some cases.

Context-aware music recommendation systems build on the principle that various personality traits connect with distinct item attributes (for example, acoustic qualities or musical genres) and that users in different emotional states and moods prefer various sorts of things [10].

Emotion-aware Recommendation systems have yielded encouraging results, demonstrating that using emotional states and reactions, recommendation algorithms may improve prediction accuracy and refine personalisation.

The emotions of users may be identified in a variety of ways. However, they all fall into one of two types. Face expressions, keystrokes, and mouse-click patterns are all examples of implicit ways of detecting emotion. Explicit methods, on the other hand, take direct input from the user. For the identification of emotion, we choose to utilise an approach based on facial expressions.

In the next sections, several Facial-Emotion Identification techniques, machine-learning models, and recommendation algorithms will be reviewed and contrasted. Further demonstration proves that the emotion-identification method used has no effect on the recommendation algorithms.

2. Related Work

The convolutional layer, the max-pooling layer, and the fully connected layer are the three different types of layers that make up a basic CNN. The feature extraction process is carried out by the first two layers, which also introduce non-linearity and feature reduction to lessen over-fitting. The last layer, referred to as a fully - connected layer, aids in classifying data using attributes that were retrieved from earlier levels. The majority of the parameters are found in the fully connected layer [1]. For the FER2-2013 dataset, the most recent model is based on CNN trained with square hinged loss [2]. Using around 5 million parameters, this model has an accuracy of 71 percent. In this design, the final fully linked layers include 98% of all parameters. Using an ensemble of CNNs, the second-best techniques in the study attained an accuracy of 66% [3].

The work in [4] discusses several deep facial recognition modules, their architectures, and loss functions. It also gives a comprehensive overview of the face-processing methods and technical challenges. Deep Facial Recognition System: Initially, faces are detected using the face detector module. These then get aligned to normalized coordinates. Finally, the Facial Recognition module is implemented. When training, different architectures and loss functions are used to extract deep features. This is then followed by feature classification.

The Multitasking Cascade Convolutional Neural Network (MTCNN) is able to beat numerous face identification tests while preserving real-time performance. It has three convolutional [5] networks: P-Net, R-Net, and O-Net. The input to the subsequent three-stage cascade network is created by taking one image and resizing it to various scales to create a pyramid of images. [5] The suggested approach is sensitive to video with poor quality as well as the distance between the faces in the video. Preprocessing using an image resolution enhancer might help improve face identification and recognition regardless of video resolution.

There was no publicly accessible dataset with emotionally annotated labels that specialized in user reviews for the task of Emotion Analysis on Reviews [6]. In [8], the authors used last.fm community tags to generate a four-cluster semantic mood space, Angry, Sad, Tender, and Happy. They compared it to current expert representations (e.g., clusters from the MIREX AMC task) and found that it was consistent, indicating that social tag folksonomies are useful for mood categorization.

Furthermore, their four clusters can be seen as representations of Russell's four quadrants in the Valence-Arousal plane.

3. Proposed System

The Proposed system consists of two phases:

3.1. Emotion Identification

3.2. Music Recommendation

3.1. Emotion Identification

Neural Networks are typically used for face classification and emotion identification. The generally used models for Emotion Identification are usually based on Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Multi-task Cascaded Neural Networks (MTCNN).

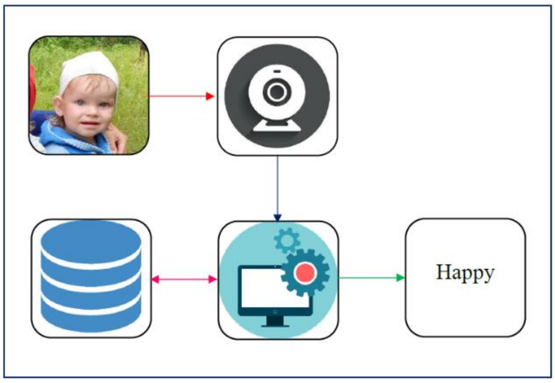

Figure 3.1: Emotion Identification Module

3.1.1. CNN for Facial Emotion Identification

Convolutional Neural Networks use the hierarchical pattern in data to build extremely sophisticated patterns from smaller and simpler patterns imprinted in its filters. As a corollary, CNNs reside on the lower end of the interconnectedness and complexity spectrum.

Compiling a model, which is effectively a neural network, is the first step in implementing a Convolutional Neural Network. The following are the parameters that must be defined when compiling a model.

Optimizers : Define Weights and Learning Rate

Losses : Choose what to minimize when training

Metrics : Function used to Judge the performance of model

Train Set : Set of Labelled Images used to train the Neural Network

Validation Set : Set of Labelled Images used to validate the Neural Network

The following settings were chosen.

Parameter	Value
Optimizer	‘adam’
Losses	‘categorical_crossentroy’
Metrics	‘accuracy’
Train Set	FER 2013 Dataset
Validation Set	AffectNet Dataset

Table 3.1.1: Arguments to Keras Model.

The Accuracy metric creates two local variables, ***total*** and ***count*** that are used to compute the frequency with which **predicted** values matches with **true** values. This frequency is ultimately returned as binary accuracy: an idempotent operation that simply divides total by count.

$$accuracy = \frac{\sum_{i=1}^{|sample|}(y_{predicted} = y_{true})}{|sample|}$$

Here,

$y_{predicted}$ = Predicted value,

y_{true} = True value,

Post Training, the model was tested against the AffectNet Dataset [7], consisting of around 49K labelled, coloured images of size 224 x 224.

3.1.2. Deep Neural Network

An artificial neural network (ANN) having numerous layers between the input and output layers is called a deep neural network (DNN). Deep Neural Networks don't loop back on themselves. As a result, they are a form of feedforward network, in which data flows from input to output.

Fortunately, a module for Facial Emotion Identification is included in Deep Face, a lightweight face recognition and facial attribute analysis framework. This module contains Deep Neural Network models that have been trained on several facial images. The Deep Face module makes detecting emotion through facial expression quite simple.

3.1.3. Multi-task Cascaded Neural Network

MTCNN, or Multi-task Cascaded Convolutional Networks, was developed as a method for both face alignment and identification. Convolutional networks at three different layers are used in the technique to identify faces and facial features such the eyes, nose, and mouth.

MTCNN has the benefit of automatically aligning poorly aligned faces. The precision may be enhanced even further this way. All the slightly misaligned photos may now be fully aligned, and a rudimentary model can be used to identify face emotions.

3.1.4. Analysing the best model for Facial Emotion Detection

After extensive research and experimentation on different neural network models such as the CNN model, MTCNN model and DNN model, the following was observed:

“The MTCNN model has the highest accuracy compared to other models. Although the train time of MTCNN is high, it can be reduced using GPUs. Hence, the MTCNN model was chosen for emotion identification.”

In the proposed system, the MTCNN model takes the user’s face as input and outputs the dominant emotion detected.

A Multi-task Cascaded Convolutional Network (MTCNN) model, trained on the FER Dataset, is used to identify the emotion. MTCNN has an added advantage of self-alignment of face compared to other models.

There are several layers of weighted nodes in the neural network. The input is processed by each layer before being sent to the one after it. During propagation, the node weights are modified. The dominating emotion is correlated to the node in the output layer with the highest weight. The model was found to perform on par with the most recent cutting-edge systems.

3.2. Music Recommendation

Content-based filtering is used to recommend music. Since there are no publicly available datasets [6] on users' emotions while listening to songs, collaborative filtering cannot be used. Hence content-based filtering approach is proposed for music recommendation [9].

The proposed approach involves identifying the emotion of each song from the MuSe dataset. The MuSe dataset consists of over 90,000 tracks along with their valence, arousal, and dominance values. Valence represents the pleasantness dimension, Arousal represents the intensity dimension, and dominance represents the control dimension. However, the dataset doesn't consist of the emotion of the track. Hence each track needs to be tagged with its corresponding emotion.

Tagging involves clustering based on VAD values. VAD values are floating-point coordinates in 3-dimensional space, which represent the overall emotion of the song. VAD values are relative and can vary based on the range of the dataset. Based on the range of the MuSe dataset, VAD values for the seven emotions are determined.

K-Means clustering is used to group songs having same emotion. Initial centroids of the 7 clusters are obtained based on the VAD range. K-Means Clustering is then performed with the initial centroids.

After clustering, we get 7 clusters, each representing one of the following seven emotions: happy, sad, angry, neutral, surprise, disgust, and fear. All the tracks with the same emotion get grouped into the corresponding cluster. Based on this, the songs in the dataset get annotated with their corresponding emotion.

It is almost impossible to determine the accuracy since it requires real-time input from users' recommendations. However, Silhouette Coefficient can be used as a metric to determine how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Silhouette Coefficient can be calculated with the following formula:

$$\frac{(a-b)}{\max\,(a,b)}$$

Here,

'a' is the mean intra-cluster distance,

'b' is the mean nearest-cluster distance for each sample.

Based on the emotion identified from the previous phase and using the tagged music dataset, top k tracks with the same emotion get recommended to the user. These get displayed as Spotify widgets, which can be played by clicking the widget.

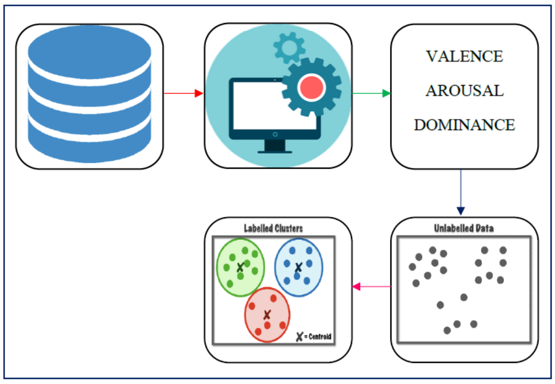

Figure 3.2: Music Recommendation Module

4. Experimentation Results

The CNN, Deep Face and FER models, which were trained on the FER-2013 Dataset, were tested against the AffectNet Dataset, consisting of around 49K labelled, coloured images of size 224 x 224. The following accuracies were reported.

Parameter	CNN	Deep Face	FER
Train Accuracy	74.83	87.35	92.19
Test Accuracy	34.2	22.8	35.8
Response time (ms)	60	97	85

Table 4.1: Compare & Contrast of Models

The following is the plot representing the train and test accuracies, and response time of each model.

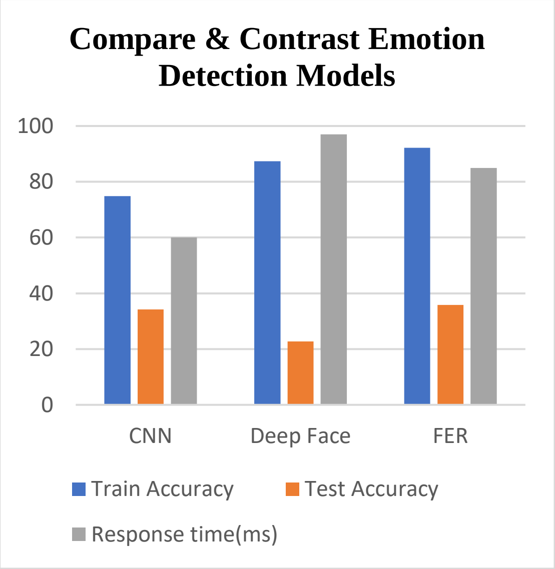

Figure 4.1: Plot for table 4.1

Table 4.1 signifies the performance of each model. MTCNN based FER Model beats the other two in both Train and Test Accuracies. It also beats Deep Face model in response time.

Furthermore, the accuracies of each model in identifying each emotion were recorded. The following shows the emotion-wise accuracies of each model.

	CNN	Deep Face	FER
Neutral	**0.326**	**0.348**	**0.574**
Sad	**0.596**	**0.19**	**0.374**
Happy	**0.92**	**0.536**	**0.69**
Anger	**0.35**	**0.178**	**0.252**
Disgust	**0.005**	**0.036**	**0.102**
Surprise	**0.116**	**0.076**	**0.184**
Fear	**0.088**	**0.232**	**0.336**

Table 4.2: Emotion-wise Accuracies of Models

The following is the plot representing the data in above table.

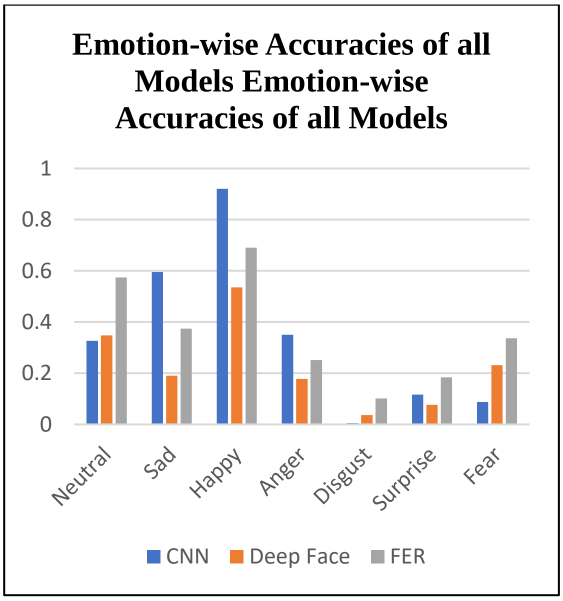

Figure 4.2: Plot for table 4.2

Table 4.2 shows the Emotion-wise accuracies of each model. It can be observed that all models fail to detect rare emotions like Disgust accurately.

It is almost impossible to determine the accuracy of Recommendation module since it requires real-time input from users' recommendations. So, K-Means clustering was done on the MuSe Dataset. The Silhouette Score and Accuracy of Clustering was calculated.

Method	Accuracy / Silhouette score
Without Initial Centroids	0.31
With Initial Centroids	0.35
Split into Train/Test	87%

Table 4.3: Silhouette Score and Accuracy for K-MEANS clustering

5. Conclusion and Future Scope

General music recommendation systems are quite ineffective because they don't consider contextual information. The aim of this work was to build a comprehensive context-aware music recommendation system, which recommends music based on the user's emotion. The overarching goal of this novel system is to blend user emotion with music recommendation, to improve accuracy and enhance the listening experience.

After extensive research and experimentation, an integrated system has been proposed that uses the MTCNN model for emotion detection and content-based filtering for music recommendation. The results suggest that the performance of the system was on par with current state-of-the-art systems.

Some shortcomings have been acknowledged in our work that opens possibilities for future work. The following shall be explored for future work:

1. Incorporate additional contextual information such as mouse click patterns, keystrokes, and user heart rate.

2. Introduce a rating system where users can rate the recommended songs. This helps in improving the accuracy of further recommendations.

3. Improve the accuracy of the MTCNN model by training against extremely large datasets using GPUs.

6. References

[1]. Lemley, J.; Bazrafkan, S.; Corcoran, P. Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision. IEEE Consum. Electron. Mag. 2017, 6, 48–56.

[2]. Yichuan Tang. Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239, 2013.

[3]. Ian Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests, 2013.

[4]. Mei Wang, Weihong Deng: Deep Face Recognition: A Survey.

[5]. A G Musikhin and S Yu Burenin 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1155 012057.

[6]. Papadopoulos Stefanos Iordanis, Athena Vakali: Emotion-Aware Music Recommendation Systems: Mitigating the Consequences of Emotional Data Sparsity.

[7]. A. Mollahosseini; B. Hasani; M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in *IEEE Transactions on Affective Computing*, 2017.

[8]. C. Laurier, M. Sordo, J. Serr_a, and P. Herrera. Music mood representations from social tags. In International Society for Music Information Retrieval (ISMIR) Conference.

[9]. H.-C. Kwon and M. Kim. Lyrics-based emotion classification using feature selection by partial syntactic analysis. 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence (ICTAI 2011), 2011.

[10]. Ji-Oh Yoo Han-Saem Park and Sung-Bae Cho. L. Wang. A context-aware MRS using fuzzy Bayesian networks with utility theory. In FSKD 2006.