Proposal: Emotional-support Music Recommender

Nghi Le
Earlham College
Richmond, Indiana, USA
nble19@earlham.edu

Abstract

Music is widely recognized as highly beneficial to human beings, particularly in relation to human emotions. On one hand, music serves as a medium for people to express their feelings through composing or performing. On the other hand, listening to music can evoke emotions, either by modulating or enhancing a person's emotional state. Regardless of the approach, listening to the right music can contribute significantly to the emotional well-being of humans. Despite that fact, current states in the art of Music Recommendation Systems (MRSs) are still under-researched when it comes to emotion-based music recommendation models. Therefore, this research focuses on developing an emotion-aware song recommender by launching a webpage, asking user to input a song they want to listen to or a song currently on listen rotation, then recognizing users' real time emotions with the help of facial emotion recognition technology CNN and lastly recommend songs by matching obtained emotions to the Music Emotion Recognition (MER) tags.

Keywords

Face detection, Emotion recognition, Convolutional Neural Network (CNN), Music recommendation systems (MRS), Content-based filtering

ACM Reference Format:
Nghi Le. 2024. Proposal: Emotional-support Music Recommender. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Although research Music Recommendation Systems has been gaining substantial interest in both academia and industry, emotion-based MRSs are still underdeveloped in recent years. This comes from the fact that current MRSs typically focus on user-item interaction and sometimes content-based descriptor, neglecting factors that significantly affect listener musical tastes and needs such as personality and emotional states, due to psychological complexities in human emotion. [17]. As a result, current MRSs often yield unsatisfactory recommendations. To build a stronger and more personalized music recommendation system, I propose an approach that takes listener emotional states into account.

2 Literature Review

2.1 Music and Emotion

The emotional state of the MRS user has a strong impact on his or her short-time musical preferences [7]. Vice versa, music has a strong influence on our emotional state. It therefore does not come as a surprise that emotion regulation was identified as one of the main reasons why people listen to music [10] [14]. As an example, people may listen to completely different musical genres or styles when they are sad in comparison with when they are happy. Indeed, prior research on music psychology discovered that people may choose the type of music which moderates their emotional condition [13]. More recent findings show that music can be mainly chosen so as to augment the emotional situation perceived by the listener [11]. In order to build emotion-aware MRS, it is therefore necessary to (i) infer the emotional state the listener is in, (ii) infer emotional concepts from the music itself, and (iii) understand how these two interrelate. These three tasks are detailed below.

Eliciting the emotional state of the listener: Similar to personality traits, the emotional state of a user can be elicited explicitly or implicitly. In the former case, the user is typically presented one of the various categorical models (emotions are described by distinct emotion words such as happiness, sadness, anger, or fear) [5] [20] or dimensional models (emotions are described by scores with respect to two or three dimensions, e.g., valence and arousal) [12].

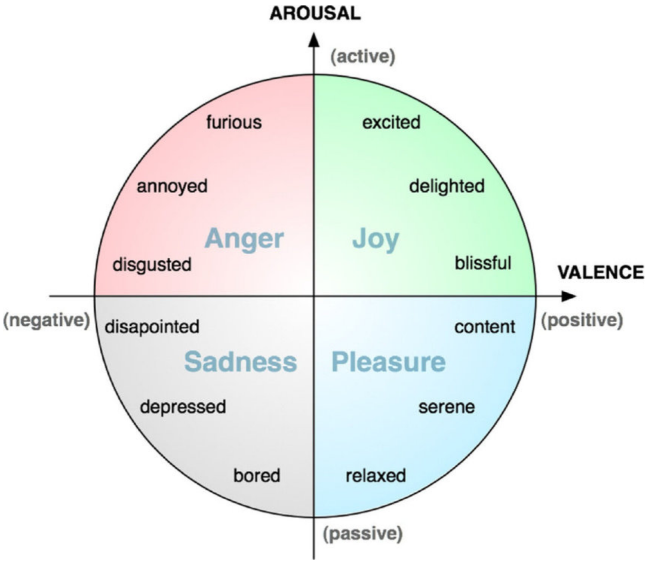

Figure 1: Circumplex model of emotion

The implicit acquisition of emotional states can be effected, for instance, by analyzing user-generated text [1], speech [3], or facial

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

expressions in video [2]. For this research, the implicit method of interest is facial expression obtained through a live camera (webcam).

Emotion tagging in music: The music piece itself can be regarded as an emotion-laden content and in turn can be described by emotion words. The task of automatically assigning such emotion words to a music piece is an active research area, often refereed to as music emotion recognition (MER), e.g., [6] [8] [19]. How to integrate such emotion terms created by MER tools into a MRS is, however, not an easy task, for several reasons. First, early MER approaches usually neglected the distinction between intended emotion, perceived emotion, and induced or felt emotion, cf. Sect. 2.1. Current MER approaches focus on perceived or induced emotions. However, musical content still contains various characteristics that affect the emotional state of the listener, such as lyrics, rhythm, and harmony, and the way how they affect the emotional state is highly subjective. This so even though research has detected a few general rules, for instance, a musical piece that is in major key is typically perceived brighter and happier than those in minor key, or a piece in rapid tempo is perceived more exciting or more tense than slow tempo ones [9].

Connecting listener emotions and music emotion tags: Current emotion-based MRSs typically consider emotional scores as contextual factors that characterize the situation the user is experiencing. Hence, the recommender systems exploit emotions in order to pre-filter the preferences of users or post-filter the generated recommendations. Unfortunately, this neglects the psychological background, in particular on the subjective and complex interrelationships between expressed, perceived, and induced emotions [15], which is of special importance in the music domain as music is known to evoke stronger emotions than, for instance, products [16]. It has also been shown that personality influences in which emotional state which kind of emotionally laden music is preferred by listeners [4]. Therefore, even if automated MER approaches would be able to accurately predict the perceived or induced emotion of a given music piece, in the absence of deep psychological listener profiles, matching emotion annotations of items and listeners may not yield satisfying recommendations. This is so because how people judge music and which kind of music they prefer depends to a large extent on their current psychological and cognitive states.

2.2 Convolutional Neural Networks (CNNs)

Shaha et al. have, in their research, stated that CNN is well-known and evidence-based to be one of the best deep learning methods for feature and information extraction [18]. CNN is a neural network architecture based on deep learning that has many practical implementations in interpreting images and visualizing data with the help of artificial intelligence. CNN is an upgraded form of artificial neural network that provides more detailed image properties for better classification. In CNN, every image (which is provided as input) is treated as a matrix. Then mathematical operations are performed over the different matrices (input image) to obtain a resultant matrix (output image) from which the required information is extracted.

Figure 2 shows a Convolutional Neural Network (ConvNet) that is composed of five layers, namely the Input Layer, Convolutional Layer, Pooling layer, fully connected layer and output layer.

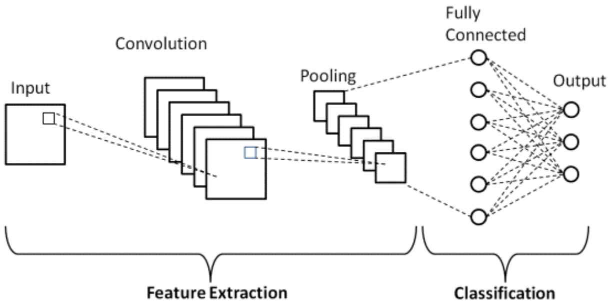

Figure 2: CNN Architecture

2.2.1. Input Layer

The input layer takes input as an image from the user converts it into a matrix, and then outputs the converted matrix to the convolutional layer.

2.2.2. Convolutional Layer

The convolutional layer is the layer present after the input layer of CNN. This is the first layer where mathematical operations are performed to extract features of the input images provided. The convolution operation is performed over the input matrix, which includes matrix multiplication of the input image and the filter (which is the MxN matrix often known as the kernel, which is used to extract properties). The result is stored in a feature map, that is developed after doing multiplication of input and kernel. Once the feature map is obtained, there are two important terminologies: padding and stride.

2.2.3. Padding

In order to preserve spatial dimensions throughout the convolution process, padding is the process of adding extra pixels surrounding the input image or feature map. It is essential to the architecture and functionality of convolutional neural networks and aids in preventing information loss at the edges.

2.2.4. Stride

While reading the image, the computer has to decide how far the filter needs to move from one position to the next across the image by stride. The filter moves across the image from the top left corner to the bottom right corner. Here, stride ensures the number of pixels (i.e., squares) that need to be skipped by the filter to read the image.

The size of the stride determines the number of features learned by the filter. The smaller size of the stride tells that more features are learned due to large data extraction. While on the other hand, more size of the stride means less features are learned due to less data extraction.

Size of feature map= (N-M+1)*(N-M+1) where,
N= number of rows in the input matrix
M= number of columns in the input matrix
NxN= size of the input matrix
MxM= size of kernel/filter

2.2.5. Pooling Layer

The pooling layer is used to compress the size of the matrix to make mathematical calculations easy, making the cost of computation less, and another is to increase the stability of ConvNet. The pooling layer contributes to a reduction in training parameters, which speeds up computation. Generally, there are three types of pooling operations i.e., max, min, and average pooling.

Max pooling chooses the maximum feature values in the selected region of the image to summarize that region. Min pooling selects the minimal feature values in the selected region of the image to summarize that region. Average pooling determines the summed value of features in a region based on its average value.

2.2.6. Fully Connected Layer

This layer multiplies the input and weight matrix and adds a bias vector to it. This layer performs the function of connecting neurons of the previous and fully connected layer. The formula for calculating a fully connected layer is shown in Equation 1.

$$y_{jk}(x) = f(\sum_{i=1}^{n_H} w_{jk}x_i + w_{j0})$$

Here,
W = weight matrix
Wo = bias vector
X= input matrix
Y= output matrix

2.2.7. Output Layer

It is the last layer of the Convolutional Neural Network, whose work is to predict the final answer by mapping the features learned from input images. The output from this layer classifies the emotion of a user, which can be any one of the emotions for which the machine is trained.

3 Preliminary Design

The preliminary design of the emotional-support music recommendation webpage will consist of two main parts: facial emotion recognition and song recommendation.

Data architecture solution will likely look something like this diagram in Figure 3.

First, create a webpage using streamlit_webrtc package in streamlit library, which is a basic web app builder. Then, face detection is conducted on live webcam video using Dlib library and OpenCV. If the face were successfully detected, the image will then be fed into the CNN model to detect emotions. If the face could not be detected, it would reopen the webcam and conduct face detection again. After successfully capturing the facial expressions and comparing to the training dataset, emotion in the live video will be extracted. Then, an analysis of the recognition of emotional states from facial expressions was performed to categorize emotions shown in each face into one of the seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral) as shown in Figure 4.

The second part deals with extracting music emotion tags from the single seed track user inputs before turning webcam on using music emotion recognition (MER) and utilize emotion data from facial expressions to recommend emotion-aware songs. The current favorite seed track will either give the system user's musical tastes to combining with the extracted facial emotions to recommend next song by using content-based descriptor filtering, or give the user's current emotional states as MER extracts emotion tags from that single track. Then, comparing to the captured facial emotion, we can obtain whether user wants to modulate (same type of emotions) or enhance (greater arousal or more positive as in the circumflex model of emotions (Figure 1)).

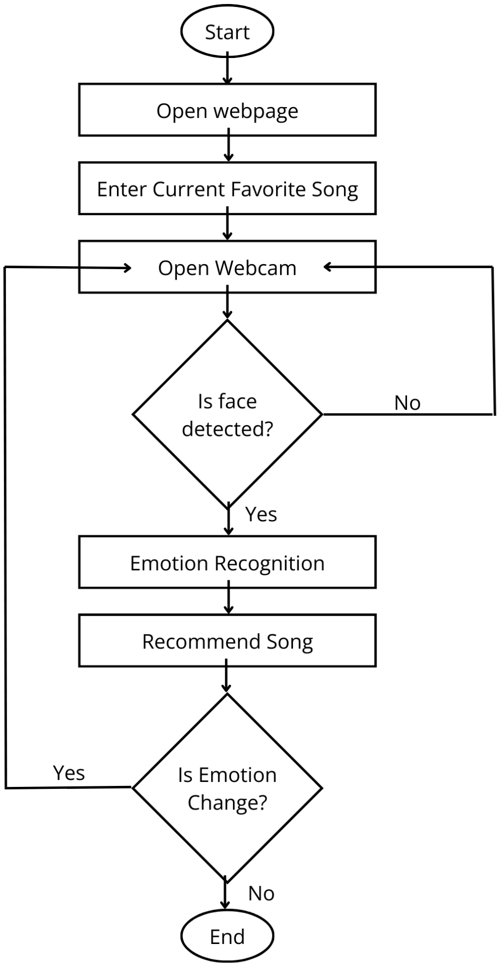

Figure 3: Flow Chart

Then, after listening to the song, if their emotion changes the process will be repeated to accommodate the new emotion recognized, or else it will terminate. With that being said, the user can continue to use the system until they completely satisfy, or are done with modulating or enhancing their emotions.

4 Evaluation plan

The dataset is self-generated with the help of an algorithm and split into two parts, training dataset and testing dataset. Each emotion is trained with 100 real-time web-captured images of the user. Because of the dataset characteristics, I intent to use the loss model and accuracy model presenting in graphs to compare the value of Epochs and Loss measure. Besides, confusion matrix will also be used. The confusion matrices are given for each emotion, in which

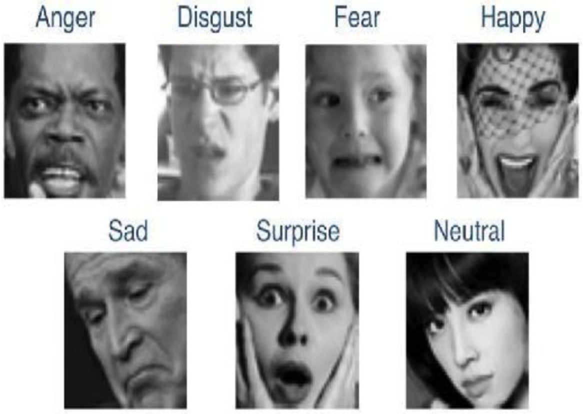

Figure 4: Seven Categories of Emotion

true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are calculated for seven emotions, i.e., Happy, Sad, Angry, Neutral, Fear, Surprise and Disgust. Diagonal elements of the confusion matrix represent the true prediction of a particular class, and the rest of the cells represent wrong predictions for that particular class.

5 Contributions

First, research in this field is still sparse due to the complexities of psychological and situational factors when integrating into music recommendation systems. Therefore, every new research or project on the subject of matter will be invaluable.

Second, emotional-support music recommender is a novel approach in hybridization models utilizing CNN and MER. While other similar approaches often neglect MER, and only use the recognized emotion (there are only 7 categories of emotion in these research) combining with an artist name input and parse the outcome straight to Youtube search engine, eg. a happy Taylor Swift song, this research steps into the unknown and complex of music emotion recognition technique and willing to take risks in exchange for new discovery.

Lastly, as mention above, instead of taking the user input as artist name, the project take "a current favorite song" of users, which is something that strikes their emotions recently or just a track in their "heavy rotation" playlist. This actually reveals a great deal about user emotional states as well as their musical tastes. By comparing this emotion to the real-time facial emotion, we are likely to predict their intention in listening to music or musical needs. For instance, if their current favorite song is a bright song but the user face is detected to have sad emotions multiple times, this is a sign that this user might want to enhance their emotions instead of modulating them.

6 Risks

As mentioned in the previous section, one of the risks is the complex of music emotion recognition tags. There is distinction between intended emotion (emotions composers or song writers had in mind when creating music), perceived emotion (emotion recognized when listening) and induced emotion (emotion felt by listeners). This problem might reduce the accuracy of the model.

The other major risk associated with this approach is that one more metric being added to the system which is the intention or purpose of user when listen to music using the single track seed. This might add too many conditions into the basic work flow as shown in figured 3. more complicated than necessary and risk the practicality and workability of the original system.

7 Special Resources

Fortunately, the project does not require any extra resources to achieve the desire outcomes. All thanks to the charge-free softwares like streamlit will be used to create a webpage, Dlib library and OpenCV for face detection and emotion recognition.

8 Timeline

Week 1: Dataset generation and collection

Create an emotion image dataset by an algorithm with each emotion being trained by 100 real-time web-captured images of the user.

Collect music dataset using the one million song dataset modification.

Week 2 + 3: Start creating the webpage and cleaning music dataset

Work with Dlib and Open CV on some basic function of the web

Cleaning and changing formats for the one million song dataset

Week 4 + 5: Get the face detection feature done and collect a MER dataset

Week 6 + 7: Train data on emotion detection and test the accuracy of the emotion recognition system

Week 8 + 9 + 10: Work on recommendation system (con-tent based filtering)

The rest: training and testing to enhance accuracy + write thesis paper and prepare for poster and presentation.

References

[1] Lily Dey, Mahim-Ul Asad, Nadia Afroz, and Rudra Pratap Deb Nath. 2014. Emotion extraction from real time chat messenger. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE, 1–5.

[2] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 467–474.

[3] Mehmet Erdal, Markus Kächele, and Friedhelm Schwenker. 2016. Emotion recognition in speech with deep learning architectures. In *Artificial Neural Networks in Pattern Recognition: 7th IAPR TC3 Workshop, ANNPR 2016, Ulm, Germany, September 28–30, 2016, Proceedings 7*. Springer, 298–311.

[4] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Personality & Emotional States: Understanding Users' Music Listening Needs.. In *UMAP Workshops*. Citeseer.

[5] Kate Hevner. 1935. Expression in music: a discussion of experimental studies and theories. *Psychological review* 42, 2 (1935), 186.

[6] Arefin Huq, Juan Pablo Bello, and Robert Rowe. 2010. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research* 39, 3 (2010), 227–244.

[7] Marius Kaminskas and Francesco Ricci. 2012. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review* 6, 2-3 (2012), 89–119.

[8] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proc. ismir*, Vol. 86. 937–952.

[9] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee. 2005. Emotion-based music recommendation by association discovery from film music.

In *Proceedings of the 13th annual ACM international conference on Multimedia*. 507–510.

[10] Adam J Lonsdale and Adrian C North. 2011. Why do we listen to music? A uses and gratifications analysis. *British journal of psychology* 102, 1 (2011), 108–134.

[11] Adrian C North and David J Hargreaves. 1996. Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition* 15, 1-2 (1996), 30.

[12] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[13] Thomas Schäfer, Fee Auerswald, Ina Kristin Bajorat, Nika Ergemlidze, Katharina Frille, Jonas Gehrigk, Anastasia Gusakova, Bernadette Kaiser, Rosi Anna Pätzold, Ana Sanahuja, et al. 2016. The effect of social feedback on music preference. *Musicae Scientiae* 20, 2 (2016), 263–268.

[14] Thomas Schäfer, Peter Sedlmeier, Christine Städtler, and David Huron. 2013. The psychological functions of music listening. *Frontiers in psychology* 4 (2013), 511.

[15] Markus Schedl, Emilia Gómez, Erika S Trent, Marko Tkalčič, Hamid Eghbal-Zadeh, and Agustin Martorell. 2017. On the interrelation between listener characteristics and the perception of emotions in classical orchestra music. *IEEE Transactions on Affective Computing* 9, 4 (2017), 507–525.

[16] Markus Schedl, Peter Knees, and Fabien Gouyon. 2017. New paths in music recommender systems research. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 392–393.

[17] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval* 7 (2018), 95–116.

[18] Manali Shaha and Meenakshi Pawar. 2018. Transfer learning for image classification. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 656–660.

[19] Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 1–30.

[20] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494.