Music Recommendation Based on Facial Expression

Mrudula K, Harsh R Jain, Amogha R Chandra, Jayanth Bhansali

B.E, Computer Science & Engineering, B.N.M Institute of Technology, Bangalore, India

***Abstract*: This paper describes various methods for music recommendation based on facial expressions. Multiple music applications suggest music based on a user's music history. But there has been some research going on for music recommendation using facial expressions/mood detection. Facial expression recognition requires image processing, which can be done using multiple algorithms, along with feature extraction, and classified into different emotions. Music genre classification requires audio processing and genres are classified based on certain audio features, and some classification algorithms. This survey paper describes and compares multiple algorithms, for the same.**

***Keywords* – Music recommendation, Facial expressions, Mood detection, Feature extraction, Image processing, Audio processing.**

I. INTRODUCTION

Music is an experience beyond words. It is a part of every known society, past and present, and is common to all human cultures across the globe. It expresses a person's feelings and mood. Everyone has their playlists defined for every mood, and certain media applications to use, like YouTube, Spotify, etc., and have witnessed song recommendations on these apps based on music history. Also, another form of non-verbal communication is facial expressions. According to a study, over 65% of human communication is non-verbal, which may include expressions, gestures, etc. Music not only heals a person's mood but also has a great impact on one's mood. Every person resorts to music, in almost every situation, and automatic detection of the user's mood is an advancement to the currently available technologies. Music recommendation systems and Facial expression detection require the use of certain Deep Learning and Machine Learning models.

Machine learning, in simple words, is training a machine to learn and improve in performance based on prior information given to it. It is a branch of artificial intelligence (AI), that makes the machine capable of learning and becoming more accurate, without the user having to explicitly program it. Machine learning algorithms read/collect data and learn at every step, for themselves.

Deep learning, on the other hand, is the branch of AI that replicates the functioning of the human brain to detect objects, speech recognition, language translation, etc. Deep learning algorithms also learn automatically from both labeled and unstructured data. Deep learning is used across all industries for several different tasks. For example, image processing/recognition in commercial applications, customer recommendation applications for open-source platforms, and research tools for medical purposes, that look for various usages of drugs for new ailments.[1]

- Recommendation systems are algorithms that focus on catering information according to particular user interests. They are mainly used for commercial purposes like advertising, e-commerce, etc. There are mainly three kinds of recommendation systems:
- Collaborative system – It uses the past interactions of the user and recommends items based on this history.
- Content-based system – It uses past interactions, along with certain other information of the user (like age, profession, location, etc.) to recommend items.
- Hybrid systems – These are combinations of collaborative and content-based systems. [2]

For the music recommendation system, given an audio files dataset, audio processing needs to be done to identify the tempo, frequency, etc. which eventually help us identify the mood or emotion of the song/audio. Generally, the librosa library in Python is used to extract features from an audio file, that help in classifying the audio genre. These features are – Temp, Frequency, Spectral contrast, and MFCC.[3] The most common Machine Learning algorithm used for music classification is K-nearest neighbors, naïve Bayesian classifiers, and clustering.[4]

Facial expression recognition has been in the field of research for more than 10 years. Training a machine to recognize expressions requires the use of Image Processing and certain deep learning models. Although many images cannot be obtained for training the model, concepts like Convolutional Neural Networks, along with image pre-processing and certain feature extraction methods, help us achieve accurate results[5]. Using these concepts, expressions like Happy, Sad, Anger, etc. can be determined by learning the features of the expression.

In this survey, different methods of recommending music based on facial expressions are analyzed.

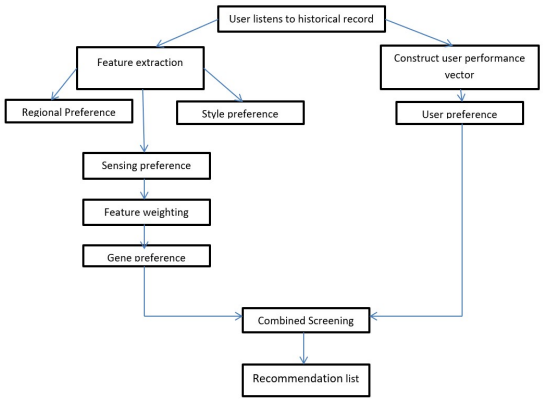
II. LITERATURE SURVEY

In [6] the authors talk about the recommendation of music to an individual based on their preference.

This recommendation system based on 2 systems:

1. Music personalized recommendation system:

- This is made of 3 parts: user preference model, music resource description, and recommendation algorithm.



-

Fig. 1 Design of hybrid recommendation algorithm

-The music resource description includes defining complexity and description includes defining complexity and abstraction level of different levels, to build a music characteristic database.

-The basic practice is to use a base gene for each type of music. Based on this gene, a collection of songs created a features record, through which feature vector is obtained. This gene is used for recommendation.

2. Key Technologies:

a. Collaborative filtering recommendation: Here, the idea is to group users into multiple groups based on similar interests. This is done by analyzing the score collected by each user for each music type.

b. Music genres: the very song is classified based on various internal and social parameters that define the characteristics of a song. These include the name of music, lyricist, gender, region, type, etc.

The architecture of the system is divided into 3 layers, namely the presentation layer, business logic layer, and data access layer.

The presentation layer is the boundary layer, used to display and is less interactive. This includes user and admin operations such as pause, play, forward, etc.

Business logic layer, responsible for data transfer and logical processing of key business of the system. The business part includes information of users and admin, users listening to songs, etc,. Logical processing adopts the above data and uses collaborative filtering and hybrid recommendation algorithms to provide a recommendation to the users.

Data access layer, mainly responsible for accessing the database.

User behaviour analysis: The fixed gene is fixed and unique for all music types. Hence, the user preference for different music can be counted. For a whole free gene, the user's annotation information can be used to get the user's preference for different musical features. This can be obtained by:

$$TP(I,j) = songTag(I,j) / sumTagCount(i)$$

Where TP(i,j) represents the preference of user i on music feature j.

The flowchart in [6, Fig. 4] shows the aforementioned system.

In [7], the authors have applied various deep learning methods to identify the emotions of the person like happiness, sadness, anger, disgust, fear, surprise, and neutral. The Kaggle Facial Expression Recognition challenge along with the FER2013 dataset was used to identify the facial expression.

The main objective of this paper is to design an effective music player which automatically identifies the person and generating a sentiment aware playlist based on the emotion of that person. The face recognition module identifies the person. The emotion module identifies the emotion of the identified person. The third module combines the results of face recognition and emotion recognition. Finally, the song is played based on that specific person's emotion. This system provides better accuracy and performance. The person is identified using CNN as shown in the [7, Fig. 1] figure below with an accuracy of 90.15%

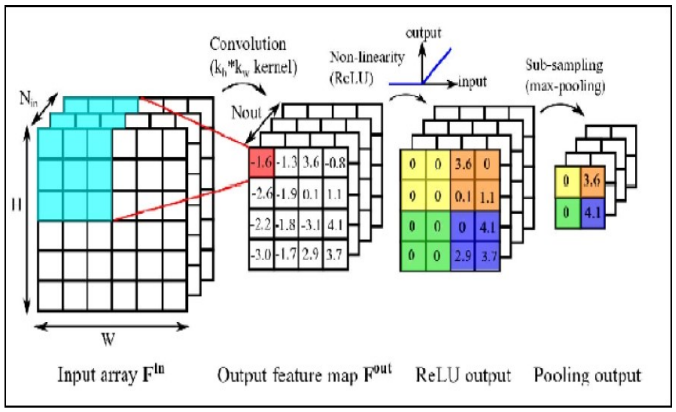


Fig. 2 Convolutional Neural Networks

For emotion detection, the Cascading classifier is used along with the FER2013 dataset. Using these emotions, the song is suggested according to the person's mood. The emotions are classified into 8 categories as shown in [7, Fig. 2] Fig. 3.

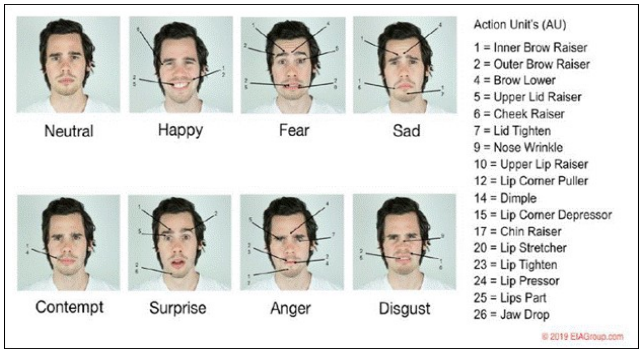


Fig. 3 Classification of Emotions

The music classification dataset comprises 300 songs based on moods. The songs are distributed in such a way that each class comprises 100 songs. Class A comprises happy and joyful songs, class B comprises of sad songs and class C comprises of anger and forcible songs, etc. The features like rhythm are extracted using MIR 1.5 Toolbox, the pitch is extracted using Chroma Toolbox, and other features like centroid, spectral flux, spectral roll-off are extracted using Auditory Toolbox. The audio signal is categorized into 8 types like joy, joy-surprise, sadness, anger, etc. Emotions extracted for the songs are stores as meta-data in the database. Mapping is performed by querying the meta-data database. Finally, the face module, emotion extraction module, and audio feature extraction module are mapped and used as an integration module as shown [7, Fig. 6] in Fig. 4.

[8] Immanuel James, J. James Arnold, J. Maria Masilla Ruban, R Saranya, and M. Tamilarasan proposed a software system based on emotion-based music recommendation.

In this paper, music is classified into 4 categories: Happy, Sad, Anger, and Surprise. This project was split into 2 phases. First, the authors developed software to capture emotion based on facial expression, and second, integrated this with web service and played the music based on extracted emotion.

The architecture of this project where facial expression is processed using a combination of 64 Action Units (AU), which is made out of multiple frames, processed from the user's video. Frames are generated using a hidden Markov model classification.

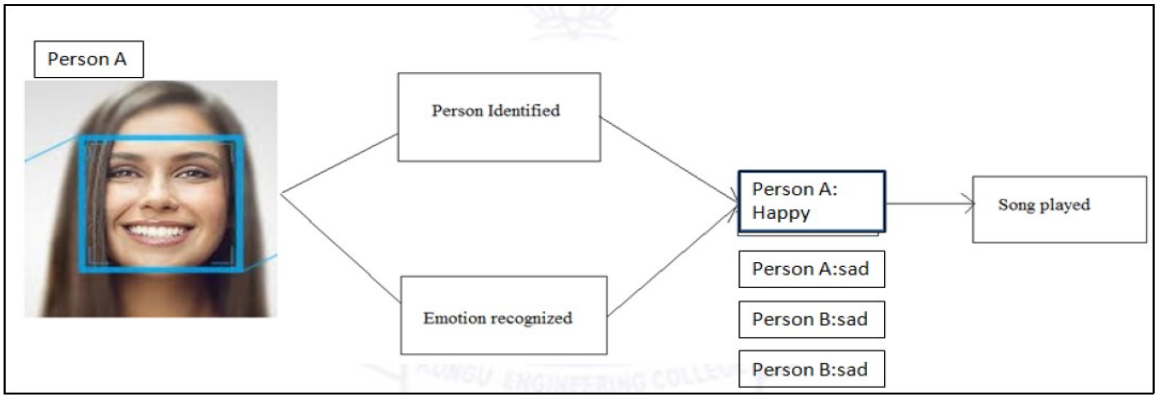


Fig. 4 Mapping of Face module, Emotion extraction module, and Audio feature extraction module

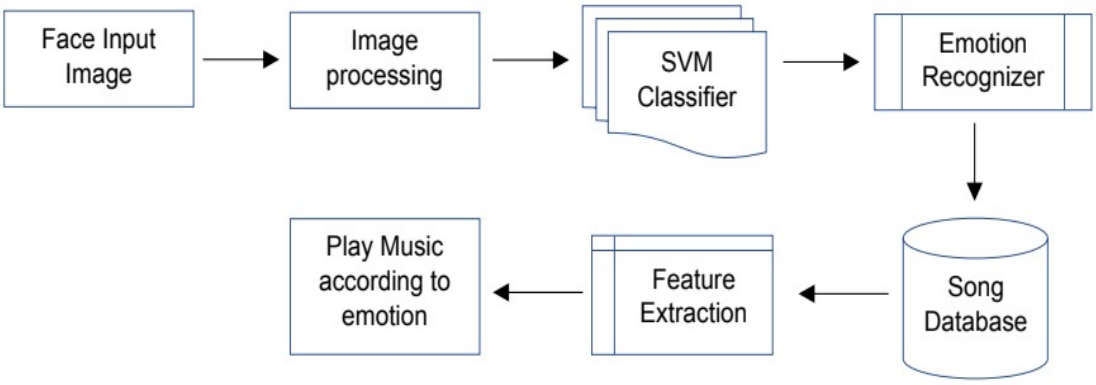


Fig 5. The sequence of a Music recommendation system for facial recognition algorithm

The system is divided into 3 steps:

1. *Face detection*:
- The image pyramid or Gaussian pyramid is used to reduce noise and other disturbances.
- Histogram of Oriented Gradients (HOG) describes the face with a set of distributions of intensity gradients.
- Linear Classifier is used to detect the face.
2. *Emotion Classification*: The extracted face is enclosed within a box and is passed to the predictor function, which extracts 68 facial landmark points to an array. The data is stored as a 68x2 array, 68 points, each with x and y coordinates.
3. *Music recommendation*: The song is played according to the emotion detected out of the four – Happy, Sad, Anger, Surprised. This is observed to obtain an accuracy of 90% - 95 %

In [9], the authors focus on building an efficient music recommendation system which determines the emotion of user using Facial Recognition techniques. This paper Proposes utilizing Support Vector Machines (SVM) as the primary characterization technique to order eight facial feelings. The block diagram [9, Graphic 1] in Fig. 5 depicts the implementation. The faces distinguished utilizing channels in OpenCV and changed over to Greyscale. The paper likewise explains robotized constant coding of outward appearances in nonstop video gushing, which is feasible for applications in which frontal perspectives can be accepted utilizing a webcam. The paper depicts utilizing Thayer's model of mind-sets to perceive the state of mind of the music piece. The edge level of a music piece is resolved and the feeling it brings is perceived via prepared neural systems.

In this paper, the authors are using OpenCV to detect the face in the image. Eigenfaces algorithm is used to recognize the face. The algorithms used for local feature extraction are Local Binary Patterns, Direct Cosines Transform, and Gabor Wavelets. The dataset that was utilized for preparing the model was Million Song Dataset given by Kaggle.

In [10], the authors have used a CNN model named Music RecNet to classify music genres and recommendations. The proposed method also detects plagiarism in music. MusicRecNet is designed to have three layers. Each layer consists of a two-dimensional convolution, an activation function (rectified linear unit), a two-dimensional maximum pooling operation, and a dropout operation. The block diagram for the proposed system is shown in [10, Fig. 3] Fig 6.

Using the dense_2 layer vector, the validation accuracy of the model goes up to 97.6% as shown in [10, Table 3] Table 1:

Table I: Comparative Classification Results Obtained by Musicrecnet

Studies	Validation accuracy %
Tzanetakis and Cook and Tzanetakis	61.0 and 79.5
Li and Tzanetkis	74.0
Holzapfel and Stylianou	63.5
Shin et al.	84.5
Elbir and Aydin	66.0
MusicRecNet	81.8
MusicRecNet + SVM	97.6

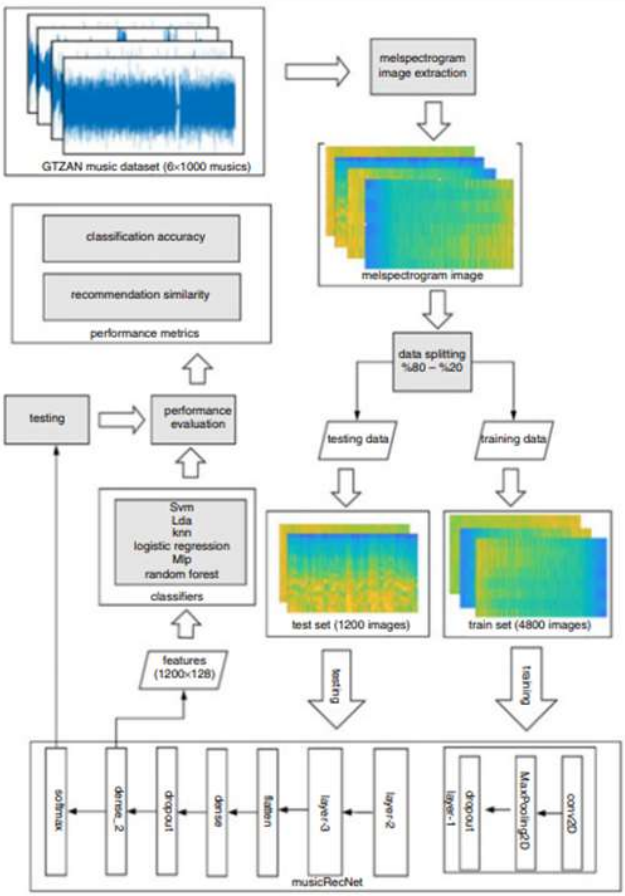


Fig.6: Block diagram of the proposed study

In [11], the proposed method, the first step is to capture the facial expression using a webcam. Then using the Viola-Jones algorithm, the face, right eye, left eye, and mouth images are retrieved. Using PCA (Principal Component Analysis) the expressions are classified as happy, neutral, sad, and surprised. In the last step, as per the result, the system provides a set of playlists that would be activated.

The result obtained when the system was tested on five random people is shown in [11, Table 1] Table 2.

Table II: Accuracy of Emotion Detection in the Proposed System

Person	Mode	Accuracy for the correct mode	Right mode
Person 1	Happy	82%	Yes
	Sad	76%	Yes
	Neutral	97%	Yes
	Surprised	69%	Yes
Person 2	Happy	79%	Yes
	Sad	73%	Yes
	Neutral	98%	Yes
	Surprised	99%	Yes
Person 3	Happy	46%	No(Neutral with accuracy 72%)
	Sad	59%	Yes
	Neutral	98%	Yes
	Surprised	52%	No(Happy with accuracy 62%)
Person 4	Happy	60%	Yes
	Sad	45%	Yes
	Neutral	84%	Yes
	Surprised	60%	No(Sad with accuracy 0.072)
Person 5	Happy	40%	Yes
	Sad	66%	Yes
	Neutral	69%	Yes
	Surprised	56%	Yes

In [12], Anuja Arora, Aastha Kaul, Vatsala Mittal, have classified music into four different genres – Happy, Sad, Angry, and Relaxed. The paper uses two different algorithms for mood recognition using facial expression and four different algorithms for mood detection. The best among the algorithms have been consequently used. For the audio data, the writers have chosen the DEAM (Database for Emotional Analysis of Music) dataset, which consists of 2800 audio files, with the average duration of the audio files being 45 seconds. For mood detection of the user, a dataset of 448.jpeg images is manually created wherein all 4 mood classes have equal images.

The feature extraction method is applied to retrieve different kinds of audio features such as harmonic features, spectral, rhythm, energy, and chroma vectors of the audio. Basic classification models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Random Forest are investigated in the paper. The audio is classified based on arousal and valence factors. Valence is positive or negative affectivity, whereas arousal measures how calming or exciting the information is. PyAudio Analysis and librosa library in the python language was used for extracting features from the audio files. The total numbers of features extracted are 36. The size of each audio file is approximately between 5mb-10mb and they are 30-60 seconds long. Audio features were put into 4 suitable dimensions namely: Dynamic, Harmony, Rhythm, and Spectral. Using Feed Forward selection it is clear that spectral, dynamic, and harmony features used together help achieve the best accuracy.

For detecting a user's mood this paper makes use of facial expressions. The pre-trained HAAR frontal-face classifier is used for detecting a user's face on the screen. Before training the model the image was pre-processed by keeping only the face portion of the image and by turning it into black and white. The fisherface algorithm was used to create the model and 16 images per mood category in 5 seconds were collected. After detecting the user's mood confusion matrix was plotted and precision has been calculated using it.

The accuracy of the audio classifier using SVM regression and Rbf classifier was 81.6% and the mood classifier using the fisherface algorithm came up to 92%.

III. CONCLUSION

To conclude, for a Music Recommendation system based on emotion recognition, it is required to first extract features from the facial recognition module. This identifies the emotion of the user. It can be done by using any of the aforementioned techniques. Next, features from the audio/music files must be extracted to identify the genre of music. The combination of the results of these two modules can be used to generate a customized playlist for the user.

REFERENCES

[1]. What is Deep Learning? [Online]. https://www.investopedia.com/terms/d/deep-learning.asp
[2]. Introduction to recommender systems. [Online]. https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada
[3]. Audio Genre Classification with Python OOP. [Online]. https://towardsdatascience.com/audio-genre-classification-with-python-oop-66119e10cd05
[4]. Introduction to Music Recommendation and Machine Learning. [Online]. https://medium.com/@briansrebrenik/introduction-to-music-recommendation-and-machine-learning-310c4841b01d
[5]. Lopes, Andre & Aguiar, Edilson & De Souza, Alberto & Oliveira-Santos, Thiago. (2016). Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. Pattern Recognition. 61.
[6]. Wu Dan "Music Personalized Recommendation System Based on Hybrid Filtration" 2019 International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS)
[7]. K. K. K. M. M. D.Keerthana, "DEEP LEARNING BASED PERSONALIZED MUSIC RECOMMENDATION SYSTEM^", IJAST, vol. 29, no. 3s, pp. 1072 - 1078, Mar. 2020.
[8]. H. Immanuel James, J. James Anto Arnold, J. Maria Masilla Ruban, M. Tamilarasn, R. Saranya "Emotion Based Music

Recommendation System" International Research Journal of Engineering and Technology (IRJET) March 2019

[9]. Samuvel, Deny and Perumal B., and Elangovan, Muthukumaran "Music Recommendation System Based on Facial Emotion Recognition" 3C Tecnologia, Glosas de innovacion aplicadas a la pymw Edicion Especial, Marzo 2020, 261-271

[10]. A. Elbir, H. Bilal Çam, M. Emre Iyican, B. Öztürk, and N. Aydin, "Music Genre Classification and Recommendation by Using Machine Learning Techniques," *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Adana, 2018, pp. 1-5

[11]. A. Alrihaili, A. Alsaedi, K. Albalawi, and L. Syed, "Music Recommender System for Users Based on Emotion Detection through Facial Features," *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*, Kazan, Russia, 2019, pp. 1014-1019

[12]. A. Arora, A. Kaul, and V. Mittal, "Mood Based Music Player," *2019 International Conference on Signal Processing and Communication (ICSC)*, NOIDA, India, 2019, pp. 333-337