Emotion aware Smart Music Recommender System using Two Level CNN

Krupa K S
Department of Information Science & Engineering
Global Academy of Technology
Bengaluru, India
krupaks@gmail.com

Ambara G
Department of Information Science & Engineering
Global Academy of Technology
Bengaluru, India
ambaragurunath@gmail.com

Kartikey Rai
Department of Information Science & Engineering
Global Academy of Technology
Bengaluru, India
kartikeyrai1824@gmail.com

Sahil Choudhury
Department of Information Science & Engineering
Global Academy of Technology
Bengaluru, India
sahilnihal7@gmail.com

Abstract— **Music plays a significant role in improving and elevating one's mood as it is one of the important source of entertainment and inspiration to move forward. Recent studies have shown that humans respond as well as react to music in a very positive manner and that music has a high impact on human's brain activity. Nowadays, people often prefer to listen to music based on their moods and interests. This work focuses on a system that suggests songs to the users, based on their state of mind. In this system, computer vision components are used to determine the user's emotion through facial expressions and chatbot interactions. Once the emotion is recognized, the system suggests a song for that emotion, saving a lot of time for a user over selecting and playing songs manually.**

Keywords— Emotion recognition, Emotion Recognition in Conversation (ERC), CNN, facial expression, semantic analysis

I. INTRODUCTION

The tremendous growth in the field of artificial intelligence and machine learning have promoted the automation of different processes which are relatively tedious and difficult to achieve manually. Emotion recognition is also one such domain gaining relevance in recent times. There are many applications where a machine can do better than human interpretation. Automated decision making systems can read a person's state of mind and thus recognize his emotional status to be applied for a variety of applications that include recommendations. It has wide applications in the area of education, entertainment, health, security, ecommerce etc [3].

Most of the time the digital music is sorted and put together based on attributes such as artist, genre, albums, language, popularity and so on. Many of the available online music streaming services recommend music based on user's preferences and his previous music listening history that employ content based and collaborative filtering recommendations. But these recommendations may not suite the current mood of the user. The manual classification of songs by learning user's preference of emotion is a time consuming task. So, recommendations can also be achieved using the physiological and emotional status of the user which are mainly captured from the user's facial expression, gestures, pulse rate, movement, speech/text interactions etc.

Several work is carried to detect emotions using facial landmarks to extract the features. Nguyen et al. [11] detected three kinds of emotions namely positive, negative and blank using 68 facial landmarks with an accuracy of 70.65%. However, the expressions of human can be understood better by applying multimodal strategy instead of single approach.

This paper work proposes a CNN based approach to recommend music by analyzing the multimodal emotional information captured by facial movements and semantic analysis of the speech/text interactions of the user, thus, intensifying the decision of the system on recognized emotions in real-time.

The organization of the paper is as follows. In section II, a brief survey of related work in the field of emotion detection is presented. In section III, the proposed system architectute is discussed followed by the hardware and software requirements in section IV. In section V, the results and performance evaluation has been discussed and finally the paper work is concluded in section VI.

II. RELATED WORK

There are different ways of emotional analysis by using facial expression, gestures, body movement, speech etc. Many research has been conducted with different approaches for detecting and classifying the physiological behavioral and emotional status expressed on the face by the users. The facial digital image is preprocessed and subjected to different algorithms for feature extraction and classification.

In 1978, Ekman and Friesen developed Action Units (AU) by using transient and permanent facial features[1]. Their work mainly focused on establishing a dependency between movement of facial muscles and expressed emotions due to variations of positive and negative emotional triggers. The framework Facial Action Coding System targeted about 44 action units on the face which were detected and measured for their intensity. Many feature-based algoriths are designed to extract human emotions as recommended by Ekman[2].

In paper [4], the author proposes a geometric based approach to recognize facial expression. Here, the features are extracted based on the movement of facial landmarks which signify the location of feature points such as eyes, eyebrows, lip corners etc. A feature vector is derived out of the distances between these feature points. Depending on the change in emotional conditions, the distances also change which is tracked with respect to the image of a neutral state. Further, the emotions are classified using SVM (Support Vector Machine) and RBFNN (Radial Basis Function Neural Network) with distance vectors as input to the classifier.

In paper [5], both geometric based approach and appearance based feature extraction using Gabor wavelet coefficients is performed using two layer perceptron.

Classification of music based on lyrical analysis is easy but not accurate. The major challenge in this approach is the language barrier that restricts classification of tokens belonging to a single language. Alternatively, emotional features and sentiment conveyed in the music can be classified using acoustic related attributes like pitch, rhythm and tempo [6]. This approach aims at extracting and defining feature vectors corresponding to characteristics of a specific mood. For example, a feature of fast tempo corresponds to an angry emotional state. So, in order to effectively recommend the songs, it is required for the system not only to know the emotions of the user but also know the mood conveyed by the song.

III. PROPOSED SYSTEM

The awareness of user emotion is the key principle behind the selection of preferred music. Fig1 shows the system architecture to arrive at the playlist based on the user's emotion. The proposed framework is a double input CNN model that uses two approaches to detect the user's emotion through facial landmarks and semantic analysis on the interactions with a chatbot as illustrated in Fig 2.

The proposed model is trained using CNN algorithm to classify the acquired facial expression. Additionally, the emotions are also captured and classified by integrating the system with an emotionally intelligent chatbot. The facial expressions captured by the webcam and the chatbot interactions are collectively used to derive the exact emotional status of the user to recommend the desired music from the predefined directories.

The emotions are recognized using a machine learning method of supervised algorithm. In machine learning, unsupervised learning models are associated with learning algorithms that analyze data used for clustering and regression analysis, thus identifying an optimal boundary between the possible outputs. So, this is done with the help of CNN.

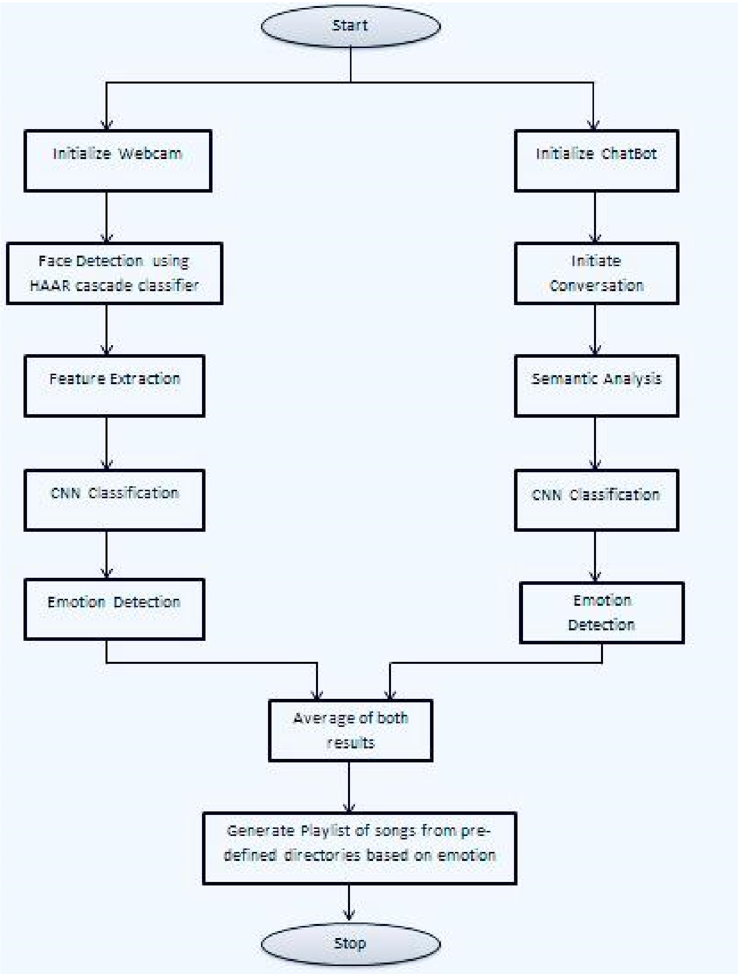

Fig 1: System Architecture

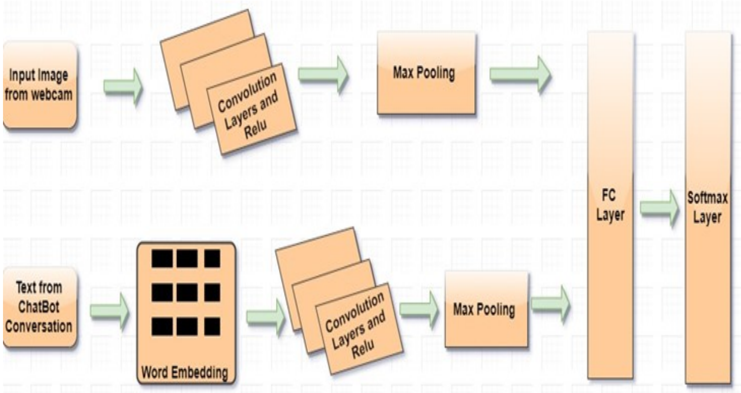
Input Image from webcam
Convolution Layers and Relu
Max Pooling
FC Layer
Softmax Layer
Text from ChatBot Conversation
Word Embedding
Convolution Layers and Relu
Max Pooling

Fig 2: CNN Model Implementation

A. Emotion detection using facial expression

Convolution Neural Networks is a widely used deep learning algorithm for image classification and recognition due to its high accuracy.

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 below:

Table 1. Emotion lables and images in 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

Out of 38,887 images, 28,709 have been used for training the model and the rest have been used for testing.

The facial image of the user is captured using a webcam. Every image is preprocessed and rescaled to an array with 48x48 grey scale values and tested with the testing data. The image is then preprocessed to detect the face using the Haar Based Cascaded Classifier inside the OpenCV framework.

Further, the model is initialized and the input image is subjected through a series of convolution layers where the high level features are extracted. The kernal/filter layer captures the spatial and temporal dependencies in the image where both the low level and high level features are identified with ReLu as the activation function.

The network is implemented with three convolutional layers to create the image feature map using 5x5 kernal size and 32, 64 and 128 filters respectively. This is followed by a 2x2 maxpooling layer to reduce the map size and computational intensity of the network. The Fully Connected layers (FC) along with Softmax function converts the resulting features into a single vector of probabilistic values to classify the emotion. Then there are two fully connected layers followed by a dropout layer to handle over fitting by 50%.

In order to speed up the performance during network construction, TFLearn library is used on top of TensorFlow, using Python. The usage of TFLearn facilitates easy and faster network construction as only the neuron layers get created, instead of every neuron. The environment also promotes easy and faster training, evaluation and classification with immediate feedback on performance of the model during training.

B. Emotion detection using semantic analysis on chatbot interaction

A chatbot is simply a robot assistant that you can engage with in a conversation to get what you need to get done. It can be designed to interpret the user queries and formulate responses just as in a natural language based conversation. The chat interface enables text messaging or voice commands to simulate human conversation[7]. Emotion recognition in conversation (ERC) is achieved using deep learning based technique. The chatbot implementation uses Google's Text-to-Speech (gTTS) and Speech-to-Text API to make the system interactive.

Chatbot begins the interaction. For loading and saving different types of audio files, Pydub along with Pyaudio module is used. Pydub is a simple tool for basic audio scripts. In the proposed framework, initially the chatbot asks the user permission to capture his facial image through a webcam. On acceptance, the captured image is preprocessed for detection of face using Haar Cascade algorithm. Further, the next level of emotions are captured through interactions. The chatbot is trained on IEMOCAP dataset [9] and modeled with text using CNN. The design of an interactive chatbot involves data set collection, preprocessing and database creation and several other stages. Detection of emotions from text conversations with chatbot involves a) Emotion Expression and b) Emotion Detection using Deep Learning approach to classify the emotion.

a. Emotion Expression: The chatbot uses the predefined responses from training dataset to generate responses for all the questions which fall within the domain. Otherwise, the chatbot uses neutral statements defined to handle the unforeseen cases.

b. Emotion Detection: Unlike rule-based approaches and non-neural network trained classifiers such as the Support Vector Machine (SVM) [10], Naive Bayes, or Decision Trees, the emotion detection and the response generation is achieved using Deep Neural Network. The primary objective is to identify and record the emotional characteristics using semantic information from text using deep learning based emotion classifier. It is important for any model to understand the context of utterance correctly before classification. Therefore, the six basic emotions of Ekman [2] namely happiness, sadness, fear, anger, disgust, and surprise are used as different categories of emotion for classification and few sample emotions are shown in Fig 3. Speech transcriptions in the form of word embeddings are fed as input to our model to detect emotion. Further, CNN is directly applied on these word embeddings without any prior information on their semantic context [8]. A vector of words from IEMOCAP dataset [9] are mapped to the context and used to train our model to predict the words in the response text. Synonymous words are grouped together in a similar context that express closer emotions, when the classes of emotions are plotted in a two dimensional space. For example, emotions like "good", "like it", 'fine', "cheerful" are classified into class of happy emotion. So, word embeddings close to each other could have the same emotion associated with it and the network can pick up on these contextual patterns during training. These embedded vectors are fed as input to our convolution layers of different sizes by varying the filter height between 12, 8, 6 and 3. Further, the max-pool layer picks up one feature from each of the different convolutional layers of the model and this pooled feature map is fed into FC (fully connected) layer to compute the class score. Finally, a softmax layer is used to perform classification.

If the score is positive then the emotion is found to be happy, and sad when the score is negative and neutral emotion otherwise. Further, the Parametric ReLU (PReLU) activation function available in tensor flow is used by the Chatbot to generate the playlist by selecting the songs from predefined directories.

ReLU typically stands for Rectified Linear Unit and is a popular activation function used in neural network. The main aim is to introduce non-linearity in the convolution layer to handle the real world data.

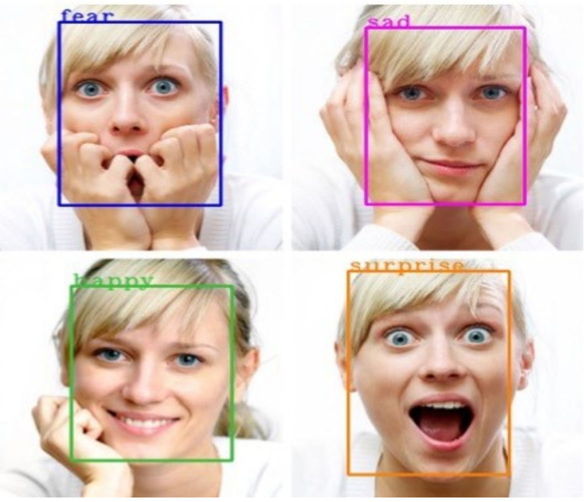

Fig 3: Sample images of few emotions

C. Music Recommendation

After detecting the emotion using proposed CNN model the next step is to recommend the music that will match the detected emotion. The desired music is picked from predefined directories for testing purposes.

Alternatively the proposed model can be integrated to access the music on web using vendor specific APIs such as Youtube API.

IV. Hardware and Software requirements

The proposed model is implemented with the below specified hardware and software environment.

Hardware:

- Processor (CPU) with 3.5 gigahertz (GHz) frequency or above (Preferably a GPU)
- A Minimum of 8 GB of RAM
- Monitor Resolution 1024 X 768 or higher
- Webcam

OperatingSystem:

- Windows 10/ Ubuntu
- Spyder/ Any IDE

Tools and Packages:

- OpenCV
- gTTS - SpeechRecognition
- Pydub, Pyaudio
- Numpy, Scikit-learn, Pandas, Wave, MatPlotLib (used for graph)

Language:

- Python 3.6

Framework:

- Tesorflow
- Keras

V. Results

In this section the accuracy and model loss has been discussed.

Loss value is an indication of the model behavior with the training and validation sets after each iteration of optimization. While the accuracy is the measure of degree of correct predictions.

Accuracy has been calculated using the Keras method. The network is trained for 60 epochs initially and the accuracy comes to be around 63% for 60 epochs. It was noticed that the accuracy seemed to increase in the last epochs. When tested for 100 epochs the accuracy came to be around 88%. In Fig 5, the Model Loss defined along x-axis is a scalar value which is minimized during training the model. So to lower the loss, closer will be the predictions to actual labels. The graph depicts loss per epoch and is calculated on training and validation sets.

The graph shows the loss calculated for 50 epochs where the training curve and the validation curve diverge at an early point indicating that the model has a good learning rate. But the convergence of the two curves shows that the model has got a good accuracy and hence 100 is chosen to be the final epochs so as to converge the curves.

Fig 6 shows the Performance Matrix where the accuracy of all the 7 emotions is represented diagonally. The X-axis represents the real emotion and Y-axis shows the emotion predicted by the proposed model. For example, in the matrix the value 0.50 (50%) for angry represents that 50% of the images in the validation set have actually been predicted as angry, which shows a right classification. But the rest, it has been wrongly misclassified as other emotions; for example, 6% as disgusted, 9% as fearful, 5% as happy etc. In the matrix, it has observed 90% of the validation set for the happy emotion that has been classified as happy. Accuracy always increases with the dataset. Since, the dataset has greater number of images for training the happy emotion, the accuracy of the emotion turns out to be good when compared to others.

Table 2. Average accuracy (%) of each detected emotion

Emotions Classified	Accuracy Rate (%)
Happy	90
Neutral	80
Surprised	77
Disgusted	62
Angry	50
Fearful	37
Sad	28

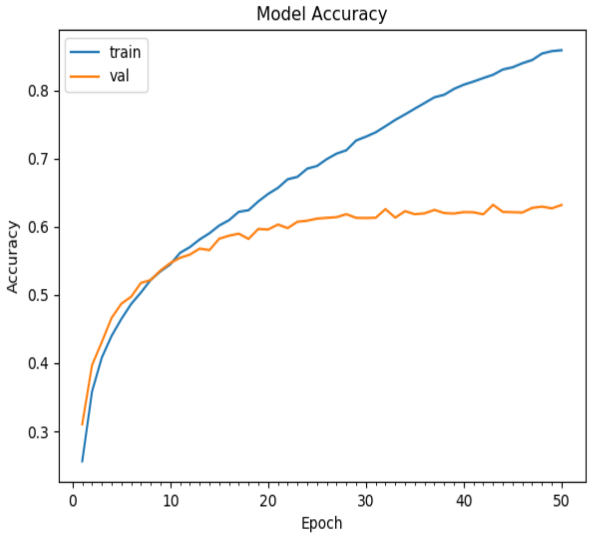

Fig 4: Accuracy of the model

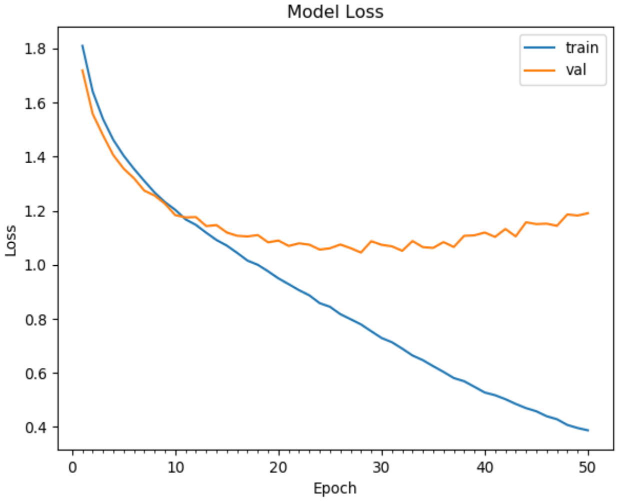

Fig 5: Model Loss

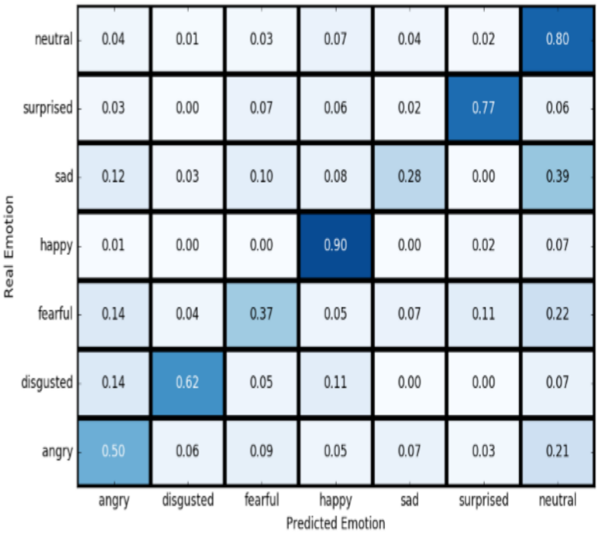

Fig 6:Performance matrix of the final model

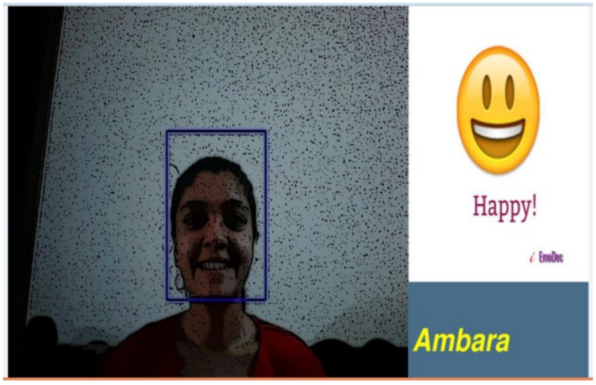

Fig 7: Output of one of the emotions from the model

VI. CONCLUSION

Music plays a major role in handling the stressful situations and emotions triggers of the user. So, it is required to recommend music that suits the current emotional needs of the user. There already exists widely used audio and video recommender systems like Spotify, Netflix, Gaana, YouTube etc which work based on search queries and not emotional needs of the user. So, the proposed CNN based model detects the emotion and generate the playlist accordingly. The model is embedded with modules for detecting facially expressed emotions and sentiments expressed with a chatbot interaction which contribute to a robust music recommender system.

As a futurer research direction of the propsoed work, the emotions detected from the proposed model can also act as input to various other emotion based use cases such as driver assisting systems, lie detector, surveillance, advertising/marketing, mood based learning, gaming etc.

REFERENCES

[1] Emanuel I. Andelin and Alina S. Rusu,"Investigation of facial micro-expressions of emotions in psychopathy - a case study of an individual in detention", 2015, Published by Elsevier Ltd.

[2] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. Emotion in the human face: Guidelines for research and an integration of findings. Elsevier 2013.

[3] F. De la Torre and J. F. Cohn, "Facial expression analysis," Vis. Anal. Hum., pp. 377–410, 2011.

[4] Bavkar, Sandeep, Rangole, Jyoti, Deshmukh,"Geometric Approach for Human Emotion Recognition using Facial Expression", International Journal of Computer Applications, 2015.

[5] Zhang, Z. Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron. International Journal of Patten Recognition and Artificial Intelligence.

[6] Remi Delbouys, Romain ´ Hennequin, Francesco Piccoli, Jimena Royo-Letelier, Manuel Moussallam. "Music mood detection based on audio

[7] nd lyrics with Deep Neural Net", 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

[8] Krittrin Chankuptarat, etal, "Emotion Based Music Player", IEEE 2019 conference.

[9] Kim, Y.: Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on EMNLP, pp. 1746–1751 (2014).

[10] Tripathi, S., Beigi, H.: Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. In: arXiv:1804.05788 (2018).

[11] Teng et al.,"Recognition of Emotion with SVMs", Lecture Notes in Computer Science, August 2006.

[12] B.T. Nguyen, M.H. Trinh, T.V. Phan, H.D. NguyenAn efficient real-time emotion detection using camera and facial landmarks , 2017 seventh international conference on information science and technology (ICIST) (2017)