International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Emophony – Face Emotion Based Music Player

Banpreet Singh Chhabra

B-tech Student, Computer Science & Engineering, Medi-Caps University, Indore 453331, India

---

Abstract - *Music plays a key role in reducing stress, building self-esteem, improving health etc. It can basically be divided into a number of different genres. People tend to select the specific music genre on the basis of their mood and interests. Hence there is really a need of a platform which automatically suggest music on the basis of emotions of an individual. Facial expressions can basically act as a form of nonverbal communication which can convey information about various moods of an individual. Hence my work, Emophony basically focusses on creation of an application to suggest songs to the end user based on their emotions and interests by capturing their face expressions. I have designed an artificially intelligent system which is capable of recognising emotions through facial expressions. Once the mood is recognized, the system then suggests a playlist of a particular genre based on the respective emotion. It will ultimately save a lot of time which otherwise is spent in searching, selecting and playing songs manually.*

***Keywords: Emotion recognition; Music Recommendation; Neural Network; Haar-Cascade; Data-set; Convolutional neural* networks; FER-2013; CNN**

1. INTRODUCTION

1.1 Music Industry

The automatic analysis, understanding and recommendation of music based on user's mood and interests, by the computer is the new possibility in the field of music retrieval. The ability to look at someone's face and guess their mood is literally the best and unique possession of human beings. If this ability is taught to an electronic device like a computer device, a humanoid robot or a mobile device, then these devices will really have an extremely valuable application in the real world. Also, Music, a most promising tool for arousing emotions and feelings, is far more powerful tool than language. Thus, listening to good music can and will always help us elevate our mood from a negative sense to a positive sense. For example, listening to upbeat songs when the person is feeling sad can help him come out of his sadness and start feeling better.

Recommending music based on user's preference of a particular music is a way to improve user listening experience. But finding the correlation between the user data like emotions and the music is a challenging task. In this work, I propose an emotion-based personalized music recommendation system to extract the correlation between the user data and the music. I have created an outstanding model for face emotion-based music recommendation having a number of applications in the new world era and will make the task of searching and listening songs even easier.

1.2 Motivation

Most of current commercial music service providers utilized metadata information such as music title, genre, album, and lyrics to search music. Some applications employ content-based methods to search for music relying on melody, rhythm, harmony etc, which provides advanced search options for different music contents. As for emotion-based methods, it is a natural and profound way to design semantic search engines of music. Although there are many works on emotion models and representations, but only a little are flexible and robust with respect to emotion-related music applications.

The earlier system focused on extracting the user's preference for a music on the basis of the user's music listening history. Based on the user's preferences, the system recommends songs to the user. But this system was limited to extracting the user's preference on the user's music listening history. However, this old system does not concentrate on how to extract the user's preference based on the user's information (such as emotions). My work will basically recommend music preference by focusing on user's emotional information.

1.3 Significance

With the outgrowth of digital music in today's era, the development of music recommendation application is helpful for users. But the existing song recommendation approaches are based on the preference of the user. However, sometimes, music recommendation according to the user's emotion is really needed. Thus, my work,

Emophony classifies the emotions of a person into several categories, by making use of convolutional neural networks.

In my work, I consider seven user emotions: happy, neutral sad, fearful, disgust, angry and surprised. Based on these emotions, the system will recommend songs to the user. For example, the system will recommend relaxing and soothing songs when a person feels sad. Party songs will be recommended by the system if a person is happy and so on.

I have used the FER-2013 data set for the task of training my model. This dataset was officially published on ICML (International Conference on Machine Learning).

After the emotion detection and classification is achieved, my platform will recommend a list of songs of a particular genre to the end user and accordingly the user can choose any song which he can play.

2. LITERATURE REVIEW

The first step towards development of a system that is able to recognize emotions through facial expression and recommend music on the basis of these emotions includes review of a number of previous researches. The publications which focusses on the way humans reveal emotions, the theory of automatic image categorization as well as the type of different genres of music and the task of music recommendation are particularly preferred for this section of work.

In the first part of this section, various necessary roles of interpreting facial expressions in emotion recognition will be discussed. It will also include surveys of previous studies on automatic image classification. The second part will focus on recommending music based on human emotions.

2.1 Emotion Recognition and Image Classification

Human emotions play a crucial role in human interaction. The facial expressions and body language are the key features in order to judge how a human interacts with others. In the nineteenth century, the English naturalist, geologist and biologist, Charles Darwin published an interesting paper on global facial expressions which played a very important role in non-verbal communication [3]. In 1971, Ekman & Friesen declared that facial behavior is universally associated with particular emotions [5]. Basically, humans and animals both tend to develop similar muscular movements corresponding to a certain mental state, despite of various other differences. Hence, if properly modelled, this similarity of face emotions can be a very convenient feature in understanding human-machine interaction. Hence, a well-trained system will easily understand emotions of any given subject.

Also, a final point to be given attention is that, emotions should not be confused with mood. Mood is basically considered to be a long-term mental state. Accordingly, mood recognition often involves longstanding analysis of someone's behaviour and expressions, and will therefore be omitted in this work.

One of the recent studies on emotion recognition by Enrique Correa, Arnoud Jonker, Michael Ozo and Rob Stolk [1] describes that the most promising concept for facial expression analysis is the use of deep convolutional neural networks. This paper basically compares different neural networks described in various other publications and uses 3 most optimal neural networks for performing the task of emotion detection and classification.

Hence, in the next section, three deep architectures in total corresponding to the paper will be subjected to an emotion classification problem. These architectures are derived from, but not necessarily equal to, the networks described at items i and ii.

(i) An outstanding publication on automatic image classification given by Krizhevsky and Hinton [7]. This work basically shows a deep neural network that resembles the functionality of the visual cortex of humans. The work uses self-developed labelled collection of 60000 images over 10 classes, called the CIFAR-10 dataset, to obtain a model which categorize objects from pictures. Another outcome of this research is the visualization of the filters in the network, such that it can be assessed how the model breaks down the pictures.

(ii) One more recent study on emotion recognition describes a neural network which is able to recognize age, gender, and emotion from pictures of faces [6]. The dataset used for this category is originated from the Facial Expression Recognition Challenge (FERC-2013). It involves a clearly organized deep network which consists of 3 convolutional layers, a fully

connected layer, and a number of small layers in between which ultimately gives an average accuracy of 67% on emotion classification, which is equal to previous state-of-the-art publications on the same dataset.

2.2 Music Recommendation

Music is a succession of tones with specific structure through time, which involves some basic perceptual attributes such as intensity, pitch, rhythm, timbre, melody, tonality, harmony. The relationship between music and emotion have been explored by a number of studies. One of the most important problems in the music psychology is how it affects emotional experience. Music has the unique ability to evoke some powerful emotional responses such as chills and thrills in listeners. Listening to music is usually an easy way to alter mood. People usually use music in their everyday lives to regulate, enhance, and diminish undesirable emotional states.

Also, musical expression of emotion is conveyed by elemental attributes of music. Many research works have already investigated the elemental attributes of music (e.g., pitch, intensity, rhythm, timbre, and tonality, etc.) contributing to the expression of emotion. For example, happy music often have relatively rapid tempos, major mode and relatively constant ranges of pitch and intensity, while sad music usually have slow tempos, minor mode and fairly constant ranges of pitch and intensity. When people listen to music, these perceptual attributes may have effect on emotion induction reflected by biophysical changes.

3. EXPERIMENTAL SETUP

3.1 Dataset

A large amount of training data can be handled with the help of the deep network and neural network. Here the performance of the model mostly dependents on the images taken for the training purpose. This type of data can help both qualitative and quantitative datasets. There is a wide range of datasets that are available for emotion recognition with high-resolution images. The one discussed is the Facial Expression Recognition Challenge (FERC-2013).

Fig1: Happy Data set

The dataset related to FERC-2013 differs from other datasets based on the quality, quantity, and cleanness of the images. There is a low resolution of 3200 in the FERC-2013 set of images.

Fig2: Sad Data set

3.2 Network

(A) The programming of networks is done by using the library of Tensorflow known as TFLearn that runs on Python. This ambiance helps to reduce the complexity of the code as there is no need to create every neuron, the similar work can be done creating a neuron layer. It can help in improving accuracy, easy to use and also provides facility to reuse the model again and again.

In this network, three convolution and two connected layers are aggregated with max-pooling layers to reduce the size of images and a dropout chance to reduce the chance of overfitting. The selection for the hyperparameter is done to bring the calculation in the convolution layer to become roughly identical. This is done to conserve the information throughout the network. Different convolution filters are used to evaluate the performance of training.

(B) The second network to test is based on previously described research by Alex Krizhersky and Sutskver, developed in 2012 that helped to classify images in about 1000 different classes for the given ImageNet

dataset. Because the model is restricted to seven emotions and due to limited computing resources, this network is considered too large.

Therefore, only used 3 convolutional layers instead of 5 and the nodes are reduced to 1024 from 4096. Also, it is seen that the original network was parallel but for the shorter version it is not necessary. Sometimes to reduce overfitting, it makes use of dropout layers.

(C) For the last experiment, the network model of Gudi is used. To start research this model could help us as [6] it also makes the use of FERC-2013 with seven emotions.

There are 48 by 48 layers in the original network in the input layer and after this input layer the model contains a convolution layer that is followed by a contrast normalization layer which is again followed by a max-pooling layer. At the end of the network there are another two-convolution layer and an output layer that in turn is connected to a softmax layer.

In this final model, the information is constantly passed by a human that is either conscious or subconscious. A human can visually interpret the information easily but, machines find it difficult. The techniques like conventional semantic facial feature recognition and various techniques are already in use for the similar purpose for physiological heuristic, but these have a lack of robustness and take more computation time. The work done by Gudi takes the help of deep learning that enables the machine to interpret the semantic feature of humans automatically rather than the manual design of the feature detector. The study related to semantic facial feature detection is analyzed by various parameters and hyperparameters. Gudi model also helps in finding the semantic features of the faces such as emotion, age gender, etc. It also gives some ideas about how to generate a 3-D actual image from the real world 2-D appearances of faces.

3.3 Evaluation

All these three networks are trained for 60 epochs with the data mentioned above. The result of this data validation is described in the research by Enrique Correa, Arnoud Jonker, Michael Ozo, and Rob Stolk. I analysed the results obtained from their research and based on their results and facts; network C seemed to be the most promising approach for emotion recognition task.

4. METHODOLOGY

4.1. Face & Emotion Detection:

The main objective of face detection technique is to identify the face in the frame by reducing the external noises and other factors. The steps involved in the FACE DETECTION PROCESS are:

4.1.1 Haar Cascade:

Firstly, I have used a classifier named, haar cascade which is used to detect the objects it is trained for. Haar Cascade is a machine learning used to detect objects in any 2-D and 3-D image or video. The approach makes the use of cascade function which super-imposes the positive images over the negative. The concept is proposed by Paul Viola and Michael [10].

4.1.2 Viola-Jones face detectors

Due to challenges faced in the field of face detection, [10] proposed a framework for object detection which used haar-like features, currently being used not only for face detection, but also for locating objects. With the help of Open CV library implementation [11], people are able to generate their own object classifiers. These classifiers are applied over an image and use haar-Features. A number of stages of the detector are created and only those image regions which pass these stages are considered as target object and are called sub-windows. The figure below shows N stages detection cascade schematic. I used detection cascading in order to remove large number of negative examples.

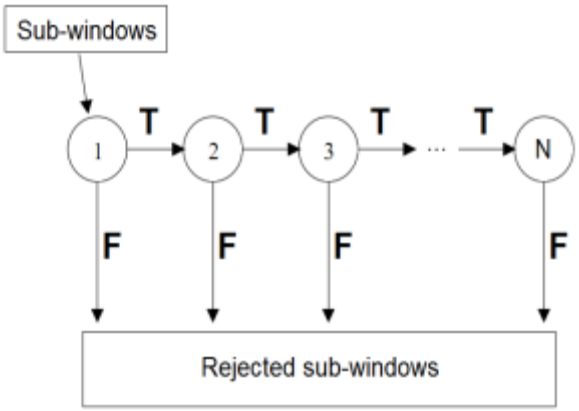

Fig. 3 Detection cascade

The Viola-Jones algorithm uses Haar-like features. The haar featured are basically, a scalar product between some Haar-like templates and the image. Mathematically,

let P and I denote a pattern and an image respectively, both of the same size N × N. The feature associated with pattern P of image I is defined by

$$\sum_{1\le i\le N}\sum_{1\le j\le N} I(i,j)1_{P(i,j)\text{ is white}} - \sum_{1\le i\le N}\sum_{1\le j\le N} I(i,j)1_{P(i,j)\text{ is black}}.$$

Since there are different lighting conditions, so in order to compensate this effect all the images should be normalized on the basis of mean and variance beforehand and those images which have a variance of lower than one, are left out of consideration because they contribute only a little.

The algorithm has four stages:

1. Haar Feature Selection
2. Creating Integral Images
3. Adaboost Training
4. Cascading Classifiers

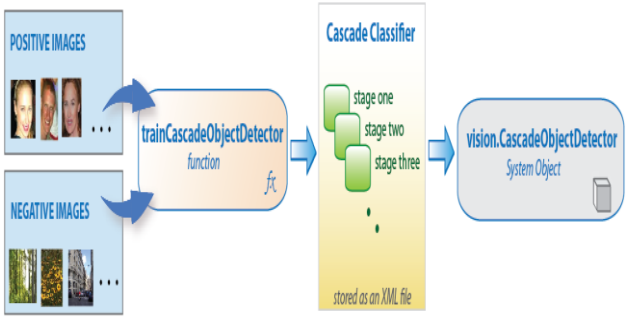

Fig 4: Working of Haar Cascade

The haar cascade model starts with the feature selection process that considers adjacent rectangular regions at a particular location in the detection window, sums up the pixel intensities in each region and calculates the difference between these sums. To make the process of feature selection fast I made use of integral images.

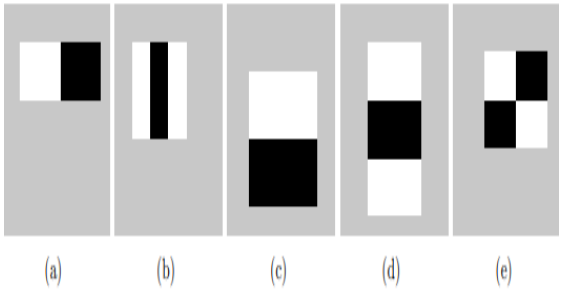

Fig 5: Haar Features

The features selected are mostly irrelevant and so I only require useful features for this purpose I made the use of AdaBoost which both selects the best features and trains the classifiers that use them.

4.1.3 Feature Selection with Adaboost

Adaboost mainly focusses on the sense of these features. The classifier will perform the mapping of an observation to a value in the finite set. For the case of face detection, the adaboost takes the form of f : R^d -> {-1,1}, }, where 1 means that there is a face and −1 that there isn't a face and d is the number of haar-features in an image.

Since the haar classifier is a weak classifier, therefore, they are organized into a cascade classifier.

After haar cascading I resized the region of images to a size of 48X48 and process the input to ConvNet which is nothing but the convolutional neural network. The output of ConvNet has enlisted the softmax scores of seven classes which turns the number into probabilities that sum up to one. As a result, the emotion corresponding to the highest softmax score is displayed on the screen.

Softmax is just like any other activation function which is used to brings some kind of non-linearity in the model data. The output of the output layer is passed to the softmax layer which converts all the values in between the range of 0-1. After conversion the sum of all the values equals to 1. Mathematically softmax function is represented as,

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \ldots, K \text{ an}$$

4.2 Music Recommendation

Now after the corresponding emotion is displayed on the screen, a corresponding list of songs will be displayed on the right side of the screen. For this I firstly created a NoSQL database which contains a whole library of songs on the basis of various genres. Now after that, I created some REST APIs for accessing that database. These APIs will retrieve various songs on the basis of corresponding emotion and will display them on interface. The end user can click on any song of his/her choice from the given options and as this action is performed, he/she will be

redirected to gaana.com web-page and the song will be played.

5. FINAL LIVE APPLICATION

As is already mentioned, live emotion recognition through video is one of the most important key-points in human-machine interaction. To show the capabilities of the obtained network, a Web-Based application is developed that can directly process webcam footage through the final model.

With use of the aforementioned OpenCV face recognition program [8], the biggest appearing face from real-time video is tracked, extracted, and scaled to usable 48x48 input. This data is then fed to the input of the neural network model, which in its turn returns the values of the output layer. These values represent the likelihood that each emotion is depicted by the user. The output with the highest value is assumed to be the current emotion of the user, and is depicted by an emoticon on the lower side of the screen.

Now after the correct emotion is displayed on the screen, a corresponding list of songs will be displayed on the right side of the screen. The end user can click on any song of his/her choice from the given options and as this action is performed, he/she will be redirected to gaana.com web-page and the song will be played.

6. CONCLUSIONS

Emophony contributes in encouraging and improving the current status of various music platforms. It will help users to play songs of their choice without even making an effort of searching. Emophony, directly recommends a list of songs for the current emotion of the user. As the user clicks on a particular song, he/she will be redirected to gaana.com web-page where the song will be played.

I have basically implemented an emotion-aware system that is designed to predict the user's dynamic emotion state through affective computing techniques.

Emophony offers a comprehensive range of number of songs, allowing users to listen a variety of songs based on their emotion efficiently, and completely. The major need of this online platform is because in real life it is really very time consuming to find appropriate song based on your current emotion.

REFERENCES

[1] Enrique Correa, Arnoud Jonker, Michael Ozo, Rob Stolk Emotion Recognition using Deep Convolutional Neural Networks, 2016

[2] Jie Deng, Emotion-based music retrieval and recommendation, Hong Kong Baptist University, 2014

[3] C. R. Darwin. The expression of the emotions in man and animals. John Murray, London, 1872.

[4] Ali Mollahosseini , David Chan, and Mohammad H. Mahoor, Going Deeper in Facial Expression Recognition using Deep Neural Networks, 2015

[5] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. Journal of personality and social psychology, 17(2):124, 1971.

[6] A. Gudi. Recognizing semantic features in faces using deep learning. arXiv preprint arXiv:1512.00743, 2015.

[7] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009.

[8] OpenSourceComputerVision. Face detection using haar cascades. URL http://docs.opencv.org/master/d7/d8b/tutorial_py_face_detection.html.

[9] R. Padilla, C. F. F. Costa Filho and M. G. F. Costa, Evaluation of Haar Cascade Classifiers Designed for Face Detection, World Academy of Science, Engineering and Technology 64, 2012

[10] P. Viola and M. Jones, "Robust real-time object detection," International Journal of Computer Vision, 2002 vol. 57, no. 2, pp. 137- 154

[11] Inte, Intel Open Source Computer Vision Library, v1. 1ore, http://sourceforge.net/projects/opencvlibrary (October 2011).

[12] Yi-Qing Wang, An Analysis of the Viola-Jones Face Detection Algorithm, Image Processing On Line, 2014

[13] Prudhvi Raj Dachapally, Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units

[14] Nima Mousavi et al. "Understanding how deep neural networks learn face expressions", International Joint Conference on Neural Networks, July 2016.

[15] Arushi Raghuvanshi and Vivek Choksi, "Facial Expression Recognition with Convolutional Neural Networks", CS231n Course Projects, Winter 2016

Snapshots of my Project:

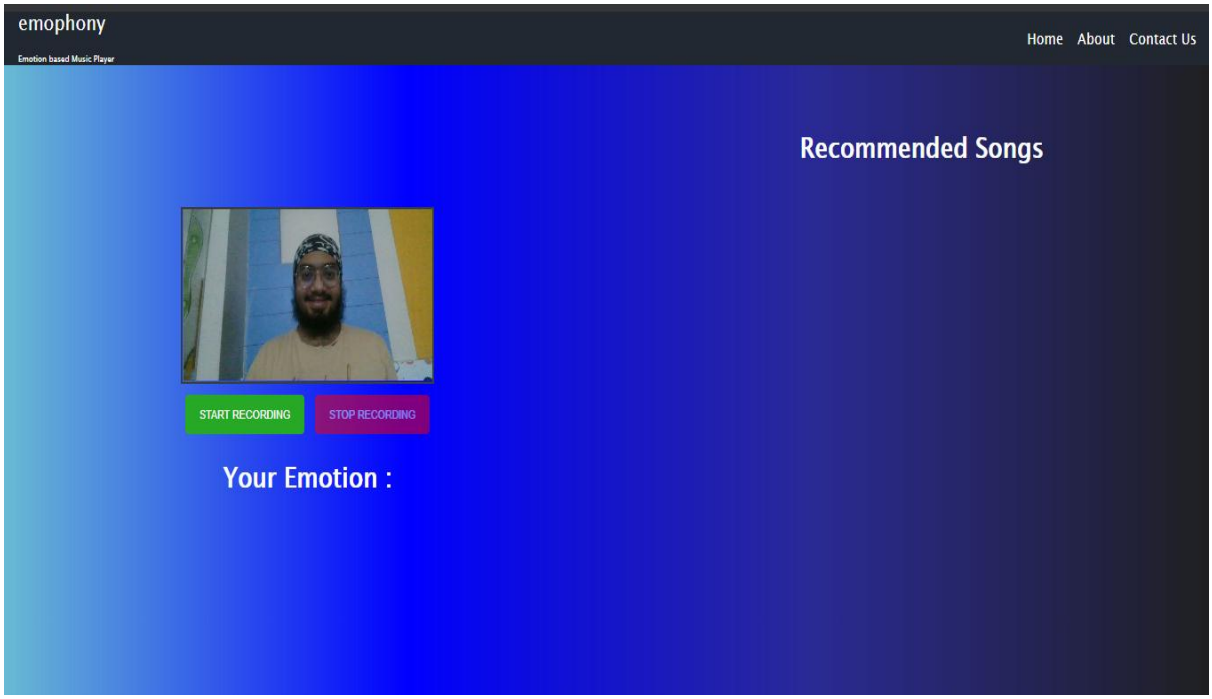

Figure 6: Snapshot 1

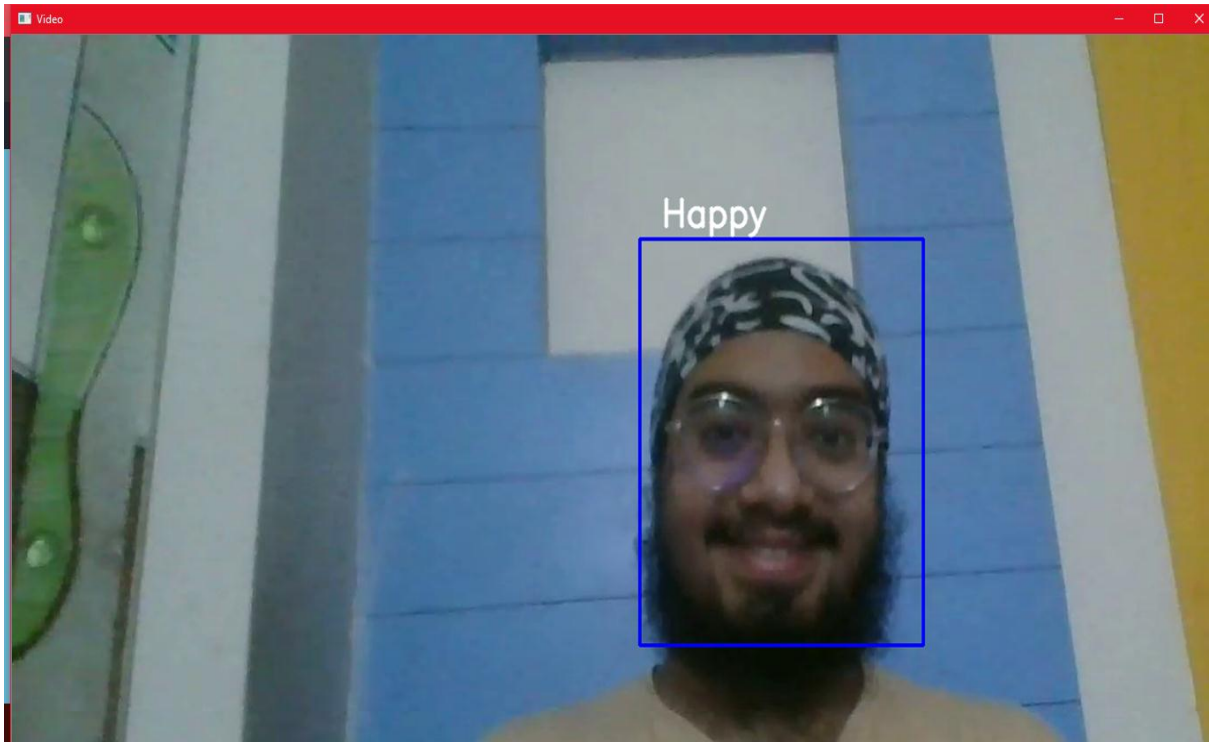

Figure 7: Snapshot 2

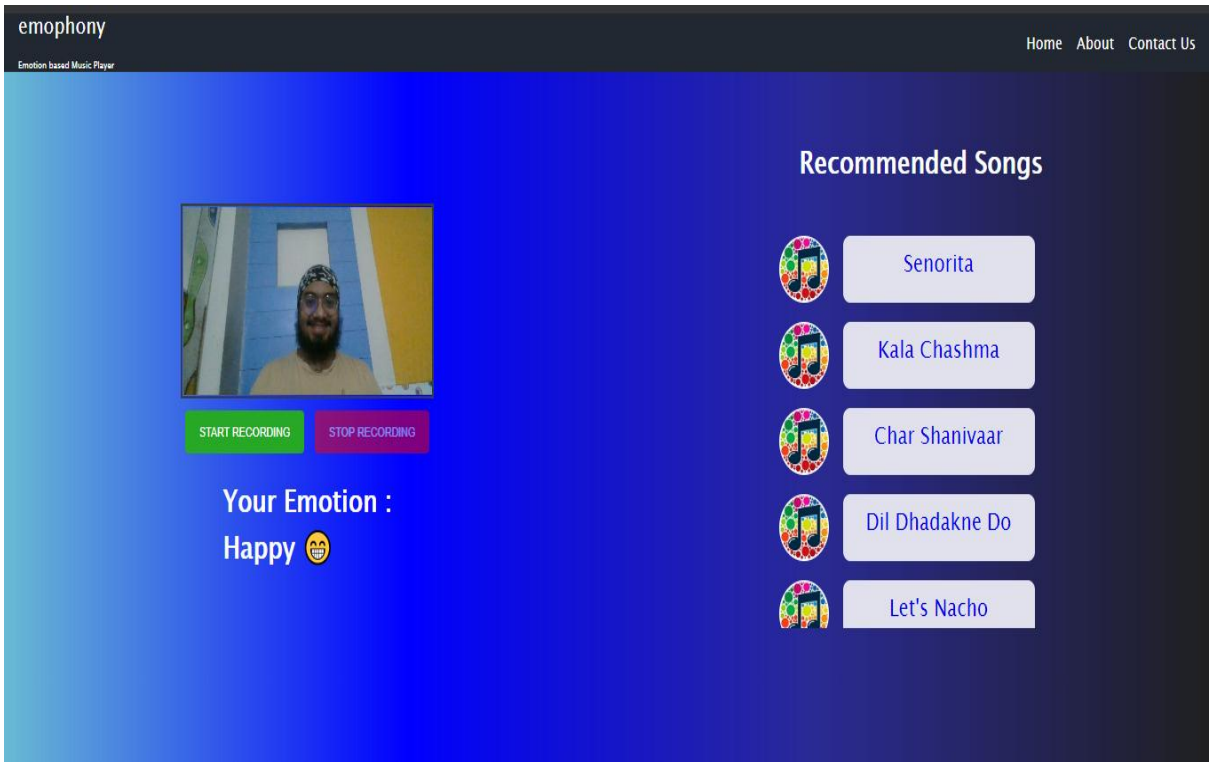

Figure 8: Snapshot 3