International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Music Recommendation System Using CNN

Amala George[1], Silpa Suneesh[2], S. Sreelakshmi[3], Tessa Elizabeth Paul[4]

UG Students, Department of Computer Engineering, College of Engineering, Chengannur, Kerala, India[1,2,3,4]

ABSTRACT: Most of the existing music recommendation systems use collaborative or content-based recommendation engines. However, the music choice of a user is not only dependent on historical preferences or music contents but also dependent on the mood of that user. In this paper, we present an effective music recommendation system, which recommends music based on the real-time mood of the user. It mainly focuses on Convolutional Neural Network (CNN) model which is based on MobileNet architecture that classifies 7 different human facial emotions. Our system consists of three modules: Emotion Module, Music Classification Module and Recommendation Module. The Emotion Module takes an image of the user's face as an input and makes use of CNN to identify their present mood. The Music Classification Module makes use of audio features to achieve a remarkable result of 98% while classifying songs into 4 different mood classes. The Recommendation Module suggests songs to the user by mapping their emotions to the mood type of the song, taking into consideration the preferences of the user. Here the model is trained, tested and validated using the manually collected image dataset with an accuracy of 98%.

KEYWORDS:Convolutional Neural Network, MobileNet, Haar Cascade Frontal face classifier.

I. INTRODUCTION

Music plays an important role in enhancing an individual's life as it is an important medium of entertainment for music lovers and listeners and sometimes even imparts a therapeutic approach. In today's world, with ever-increasing advancements in the field of multimedia and technology, various music players have been developed with features like fast-forward, reverse, variable playback speed, local playback, streaming playback with the multicast streams and including volume modulation, genre classification etc. Although these features satisfy the user's basic requirements, yet the user has to face the task of manually browsing through the playlist of songs and select songs based on his current mood and behaviour. That is the requirements of an individual, a user sporadically suffered through the needs and the desire of browsing through his playlist, according to his mood and emotions. Using traditional music players, a user had to manually browse through his playlist and select songs that would match with his mood. This task was a labour intensive and an individual often faced the dilemma of landing at an appropriate list of songs.

In recent years, with the popularization of deep learning, great progress has been made in image classification. Convolutional neural networks are one of the most popular deep learning architectures for image classification, recognition, and segmentation. Convolutional neural networks are built like a human brain with artificial neurons and consist of hierarchical multiply hidden layers. These artificial neurons take input from the image, multiply weight, add bias and then apply activation function. So, artificial neurons can be used in image classification, recognition, and segmentation by performing simple convolutions. By feeding the convolutional neural network with more data, a better and highly accurate deep learning model can be achieved.

Deep learning-based facial expression recognition is one of these methods to detect emotion state (e.g., anger, fear, neutral, happiness, disgust, sadness and surprise) of human. This method aims to detect facial expressions automatically to identify the emotional state with high accuracy. In this method, labelled facial images from facial expression dataset are sent to CNN and CNN is trained by these images. Then, the proposed CNN model makes a determination which facial expression is performed.

II. RELATED WORK

Various methodologies have been proposed to classify the behavioural and emotional state of the user. Mase et al. focused on using movements of facial muscles while Tian et al. [1] attempted to recognize Actions Units.

G.Yang et al [2] proposed a DNN model which uses vectorized facial features as input. The model can predict different emotions with an accuracy of 84.33%.

Liu et al [6] use the fer2013 dataset and two-layer CNN to classify different facial emotions. They have also compared it with 4 different existing models and the proposed model yields a test accuracy of 49.8%.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Determination of the emotion expressed by the audio piecehas been performed based on the Arousal-Valence model ofemotion proposed by Thayer [3]. This model suggests that the emotion of a song is accurately expressed as a combination of two factors - the arousal, or energy of the song, and the valence, or stress of the same piece. A high valence indicates a positive emotion, while a low value represents a negative Vibe.

Another method for music mood classification is using acoustic features like tempo, pitch and rhythm to identify the sentiment conveyed by the song. This method involves extracting a set of features and using those feature vectors to find patterns characteristic to a specific mood [4], [5].

.

III. METHODOLOGY

In this paper, we proposed a music recommendation system based on CNN (MobileNet architecture) for facial expression detection to recommend corresponding songs based on the user's mood. The whole system is fragmented into three modules: Emotion Module, Music Classification Module and Recommendation Module.

1. EMOTION MODULE

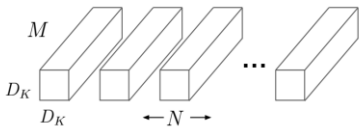
This module deals with the usage of the convolutional neural network which uses a MobileNet architecture and Haar Cascade Frontal face classifier which is used to detect the presence of face and extract the features needed to form a pattern that matches the facial expression. The CNN is trained manually for the classification of 7 emotion states (happy, sad, surprise, angry, disgust, fear, neutral) with an accuracy of 98%.

a) MobileNet Architecture

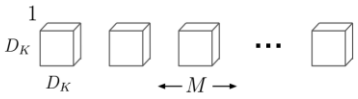
The core layer of MobileNet is depthwise separable filters, named as Depthwise Separable Convolution. The network structure is another factor to boost the performance. Finally, the width and resolution can be tuned to the trade-off between latency and accuracy. Depthwise separable convolutions which are a form of factorized convolutions which factorize a standard convolution into a depthwise convolution and a 1×11×1 convolution called a pointwise convolution. In MobileNet, the depthwise convolution applies a single filter to each input channel. The pointwise convolution then applies a 1×11×1 convolution to combine the outputs the depthwise convolution.

Table 1. MobileNet Body Architecture

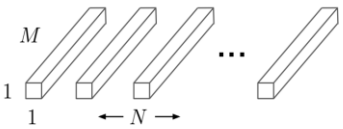
Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
5× Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters



(c) 1 × 1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

b) Dataset Details

The dataset for the proposed model includes 7 different facial emotions (happy, sad, surprise, angry, disgust, fear, neutral). These are collected manually using a 48 MP camera. Each image has a pixel size of 1920 x 2560. The dataset split is shown in Table 2. Each class consists of the same number of training samples so that they are not biased. There is a total of 21000 images corresponding to these emotions. The breakdown of the images is as follows: happy with 3000samples, sad with3500 samples, neutral with1500 samples, angry with 4000 samples, surprise with 2000 samples, disgust with 2500 samples and fear with 2500 samples.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Table 2: Dataset Details

Number of classes	7
Number of training images	21000
Number of validation images	2600
Number of testing images	2600

c) Model Description

A multi-layered convolutional neural network is programmed to evaluate the features of the user image. The convolutional neural network contains an input layer, convolutional layers, ReLU layers, pooling layers, fully-connected layers and an output layer.

Input layer: The input layer has predetermined dimensions. So, for pre-processing the image, we used OpenCV for face detection in the image before feeding the image into the layer. Haar Cascade Frontal face classifier which is used to detect the presence of face and crop the required portion of the face. The cropped face is then fed into the input layer as a NumPy array.

Convolution layer: The matrix image with pixel values is entered into it. Next, the software selects a smaller matrix called a filter. This the filter moves along the input image. The filter's task is to multiply its values by the original pixel values. All these multiplications are summed up and a number is obtained in the end. Since the filter has read the image only in the upper left corner, it moves right by 1 unit performing a similar operation. After passing the filter across all positions, a matrix is obtained, but smaller than an input matrix.

Relu Layer: A nonlinear layer is added after each convolution operation. It has an activation function, which brings nonlinear property.

Pooling layer: It performs a downsampling operation on the image to reduce its size. After completion of the series of convolutional, nonlinear and pooling layers, it is necessary to attach a fully connected layer.

Fully connected layer: This layer takes the output information from convolutional networks. Attaching a fully connected layer to the end of the network results in an N-dimensional vector, where N is the number of classes from which the model selects the desired class.

Output layer: The output is represented as a probability distribution for each emotion class.

2. MUSIC CLASSIFICATION MODULE

This module deals with the procedure that was used to identify the mapping of each song with its mood. Based on these features, we trained an artificial neural network which successfully classifies the songs in 7 classes with an accuracy of 98.65%. The music dataset comprises of 500 songs spread across seven moods. The distribution of the songs is as follows: class A with 100 songs, class B with 93 songs, class C with 100 songs, class D with 97 songs, class E with 150 songs, class F with 135 songs and class G with 140 songs. The songs were manually labelled. All the features were extracted using Python 3.6and relevant packages

3. RECOMMENDATION MODULE

This module is responsible for generating a playlist of relevant songs for the user. It allows the user to modify the playlist based on her/his preferences and modify theclass labels of the songs as well. The working of this module is shown in figure 1 and 2 as training and testing phases. First, the input is acquired in real-time so the camera is used to capture the video and then converted into image frames. The frames that are obtained are considered in all frames and all pixel formats for the purpose of emotion classification. Haar Cascade Frontal face classifier which is used to detect the presence of face and extract the features needed to form a pattern that matches the facial expression. The efficiency of the classifier is about 90-95%. So that even when there are any changes in the face due to environmental conditions the system can still identify the face and the emotion being expressed. The emotions are then identified using CNN using the training set which has been already provided to the system.After identifying the emotion, the corresponding class label is assigned to the identified emotion. As we can see that the classified songs are mapped with the user's mood. So that after getting the user's mood the system will automatically play a random song from the corresponding music class which is based on the emotion detected.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

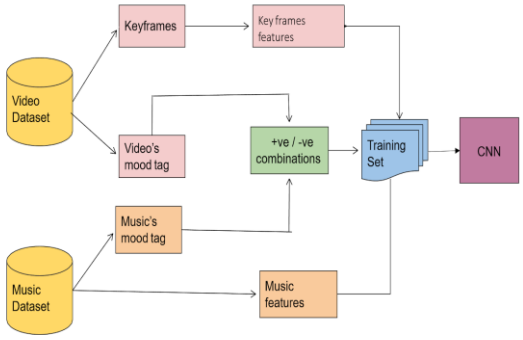


Figure1: Training phase

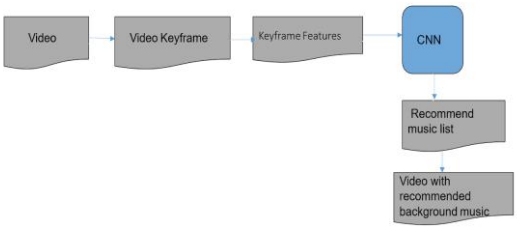


Figure 2: Testing phase

IV. RESULT

In this study, *Keras* and *TensorFlow* libraries were used for training MobileNet CNN architecture and prediction of emotion states with proposed deep learning model. According to experiment results, training loss was found 0.0887; training accuracy was found 98.43%; validation loss was found 0.2725 and validation accuracy was found 98.81%.

V. CONCLUSION

In this project, we presented a music recommendation system based on emotion detected. The system uses a two-layer convolution network model for facial emotion recognition. The model classifies 7 different facial emotions from the image dataset. The model has comparable training accuracy and validation accuracy which convey that the model is having the best fit and is generalized to the data.

We also recognize the room for improvement. It would be interesting to analyze how the system performs when additional emotions are taken into consideration. User preferences can be collected to improve the overall system using collaborative filtering. We plan to address these issues in future work.

REFERENCES

[1] Ying-li Tian, T. Kanade, and J. Cohn, "Recognizing lower. Face action units for facial expression analysis," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Mar. 2000, pp. 484–490.

[2] G. Yang, J. S Saumell, J Sannie, " Emotion Recognition using Deep Neural Network with Vectorized Facial Feature" 2018 IEEE International Conference on Electro/Information Technology (EIT).

[3] Thayer, R. E. *The Biopsychology of Mood and Arousa*l. Oxford University Press, USA, 1989.

[4] Bo Shao, Dingding Wang, Tao Li, and Mitsunori Ogihara, "Music recommendation based on acoustic features and user access patterns, "IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 8, Nov. 2009.

[5] A. S. Bhat, V. S. Amith, N. S. Prasad, and D. M. Mohan, "An efficient classification algorithm for music mood detection in western and Hindi music using audio feature extraction," *in 2014 Fifth International Conference on Signal and Image Processing, Jeju Island*, 2014, pp. 359–364

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Lu Lingling Liu, "Human Face Expression Recognition Based on Deep Learning-Deep Convolutional Neural Network", 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA).