Proceedings of the Sixth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022)
IEEE Xplore Part Number: CFP22J88-ART; ISBN: 978-1-6654-8271-4

2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-8271-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECA55336.2022.10009391

CNN based Music Recommendation system based on Age, Gender and Emotion

Jayadeep Jayakumar
Dept. of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.p2ebs20010@cb.students.amrita.edu

DR. Supriya P
Dept. of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
p_supriya@cb.amrita.edu

***Abstract*—Music is a soothing combination of sounds that can heal a person mentally and emotionally. Listening to music can improve their mood and productivity, as it helps to de-stress/reduce anxiety and enhance sleep quality. Here, the proposed music recommendation system is based on Age, Gender, and Emotion. For a better-personalized recommendation system, the age and gender should also be considered in addition to the emotion, further depending on different age groups and gender, the preference also changes. The entire system can then be divided into three neural network-based models to detect age, gender, and emotion, respectively, and depending on this combination, the personalized playlist has been suggested.**

***Index Terms*—Convolution Neural Networks, Facial emotion detection, Music Recommendation System, Music playlist**

I. INTRODUCTION

As the human clan had to go through the covid pandemic, everyone's mental health was hit because of the lockdown and reduced access to the health service sector. Various studies show that the number of patients admitted due to depression and anxiety related to Covid-19 has increased worldwide. Listening to music can be a source of mental peace to all. The music consists of multiple elements such as Rhythm, Harmony, Pitch, Tempo, etc. A combination of these elements helps in changing human mood to an extent. Music has no barriers; it can travel through any region or language without any issue. Studies show that human body produces more Oxytocin and Serotonin when one listens to songs. These are known as Instant mood boosters; that's why most of human beings can feel mood changes when listening to one's favorite music.

Recommendation systems for music have been around a long time. Nonetheless, in most cases, the recommendation is based on past played songs or similarities between other users. In this paper, the suggested method is based on the person's current mood, age group, and gender. The Emotion detection module classifies the mood into happy, sad, or neutral, while the Age Classifier module detects age according to different age groups such as kids, teens, adults, etc. The upcoming sections explains the details about the literature survey carried, proposed system, datasets used, results and conclusion. These are explained in detail in Section two, three, four and five respectively.

II. LITERATURE SURVEY

Facial expression recognition has been a field of study for many years. Emotion detection can be used in different areas such as medical services, Marketing research, Video gaming, etc. So, it's essential to find a model which can predict emotions with at most accuracy. There are many difficulties in getting accurate emotion detection starting from a titled image, or less illumination of images etc. In [1], a graph CNN (GNN) system that can predict the facial expression. The input from the camera is resized to 128x128, and using landmark detection, the features are extracted and then given to a GNN. The proposed method was performed on the FER2013 dataset and yielded an accuracy of 95.85%. According to the trials conducted in [2], two methods were carried out to find the facial emotion. The first method uses Linear Discriminant Analysis (LDA), and the second is Facial landmark Detection. The images are cropped, converted to grayscale, and given to the two systems. The facial landmark detection method performed better with an accuracy of 84.5% compared to the 73.9% using LDA. The facial expressions are mapped onto the six hexagonal edges, and by connecting these six hexagonal to the six expressions, the changes occur when one person is changing their expression. From this model, it's found that each expression changes at least one part of the hexagon, e.g., Fear causes a change in the upper triangle portion of the hexagon [3]. Dong Yoon Choir et al. discussed the method to find fine micro-expressions by two definitional landmark feature extraction [4]. The method uses existing landmark information into 2D image information and gives this 2D landmark feature into a convolution neural network. The proposed system got an accuracy of 77% with the CK+ image dataset. Automated Emotion detection using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) is proposed in [5]. The HOG extracts and the facial and extracted features are fed to the SVM classifier for recognition of the exact emotion. Using the JAFFE and CK database, the system yields an accuracy of 97% and 98.6%, respectively.

Age and gender are other important attributes that can be estimated from a single image for essential applications. A CNN-based method using the RoR, ImageNet is used to pretrain the model. The resulting model performed better than most CNN methods for age and gender estimation [6]. Using the time-series data from the mobile application and feeding it to a deep convolution neural network model for the classification of age and gender [7]. The method was tested on the Intage Single Source Panel obtained from INTAGE obtained an accuracy of 61% for gender classification and 38% for age range classification. In [8], the video from the CCTV video surveillance system to calculate the age and gender of the people walking by using the Wide Residual Networks. Another method is implemented by detecting the density of wrinkles [9]. The image taken from the database is converted to an HSV image and, by using edge detection, creates binary masks for the given image. According to the region of interest, the wrinkle density is calculated for different age groups of people. In [10], the proposed system uses pre-trained VGG19 and VGGFace to detect the gender and calculate the age of the person from the images. The gender classification yielded an accuracy of 98.7%.

Various research is happening on Music recommendation systems based on different attributes such as Emotion, ECG, HRV and also depending on past played song lists. The Emotion from the user is taken from an intelligent wearable watch which collects the galvanic skin response and Photo plethysmography [11]. This data is fed to build content-based recommendation systems already. The valence and arousal value can be taken from the wearable and calculates the Emotion of the person. Also, various Emotion-based system is studied using deep learning algorithms. In [12], emotion detection using a convolution neural network is used for the music recommendation system. The author uses two CNN models: One for emotion detection and the second for Music recommendation. Similarly, an Emotion-based music recommendation considers the four emotions: Happy, Angry, Sad and Calm, using CNN implemented in [13]. The emotion classification is based on Recurrent Neural networks and Convolution Neural networks using the FER2013 dataset, and the Amrita custom dataset is used to train the images and voice for the classification of emotion [14]. A system with meta-learning using a Siamese network is used for the classification of the emotions, which gives better accuracy with a small sample of data for different poses [15]. The classification of the collection of songs using signal processing and artificial intelligence for the deployment of a music recommendation system with Tamil songs in [16]. In [17], the spectral features of the music is extracted, classified into four emotions and the accuracy is tested for SVM and ANN.

III. PROPOSED SYSTEM

The Fig 1. shows the process flow chart of the proposed model. The system is divided into three modules Emotion detector , Age Estimator and Gender detector. From the input image the face is detected and the region of face is converted to grayscale. Then the cropped area is fed to the three different modules. The output from each module combined opens the required playlist according to the mood, age and gender.

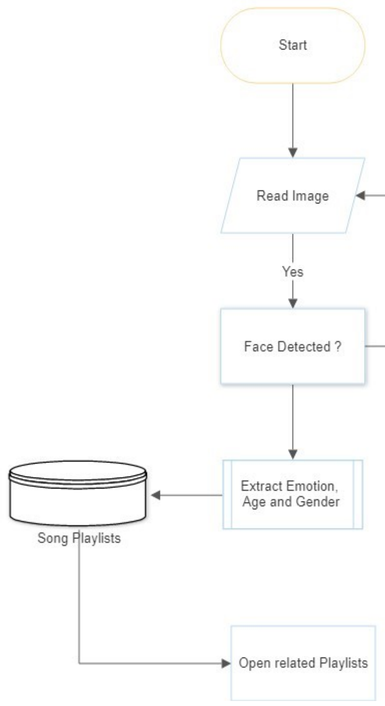


Fig. 1. Process flowchart

A. Emotion Detection Module

a) Dataset: The Facial Expression Challenge or the commonly known as FER2013 dataset and the CK+ dataset is used for the training of the Emotion detection module. The FER2013 contains 7 emotions: Angry, Fear, Disgust, Sad, Happy, Neutral, Surprise). The total number of images is 26,217. Each image is grayscale and cropped to 48x48 pixel. The proposed study focuses on 3 emotions: Happy, Sad, and Neutral. The total images from these three categories are 21264 images. The images are split into 80:20 for the train and test set. Another Dataset used for Emotion recognition is the extended Cohn-Kanade or the CK+ dataset. This dataset consists of images of people from the age 18 to 50 from different places. It actually is video of people having showing different emotions. Total 981 image samples are available in the dataset with different emotions combined.

b) CNN model Description: The CNN model used for emotion detection consists of different layers of convolution layers, pooling layers, ReLU layers and connected lay-ers and an output layer. The input layer gets the image from haar-casacade face detector, cropped and converted to grayscale. The image is converted into image arrays using NumPy. Next comes the convolutional layer uses different detectors which moves through the image to find the exact features. The each filter have size of 3x3 which goes through the image array. ReLU is used as the activation function. After every three sets of convolution layer MaxPooling is used for dimensionality reduction and the size of this averaging filer is 2x2.A flattening layer is used to reshape to match with the exact number available in the tensor. The final dense layer connects the actual input image to the different class what we specify ie Happy , Sad or Neutral. Sigmoid activation function is used after the dense layer for the output layer.

B. *Gender Detection Module*

a) Dataset: Using the UTK faces dataset the Gender Detection module is trained and validated. The Dataset contains 23,708 images labelled with age, gender, ethnicity. The images are cropped to 200x200 size. The images are of different expression, lighting etc. The age of people varies from 1 to116 years.

b) CNN model Description: The image input from the UTK dataset is resized to 100x100 and before it goes to the input layer of CNN the image is transformed to grayscale. After the input layer comes the convolution layer with filters which moves through the entire image and pooling layer is used to match the dimensionality and last will be the fully connected layer. The number and the combination vary according to the architecture. Each of the convolution layer gets some number of filters and here it starts with 32. Usually, the size of the filters can go up to 7x7 but since out input image size is less than 128x128 we use 3x3 for better results. The stride is set at (1,1) for moving along the two-dimensional axis respectively. Stride (1,1) moves one pixel through each axis. The activation layer of each convolution network is ReLU. After every convolution layer the maxpooling2d is used.

C. *Age Estimation module*

a) Dataset: Age estimation just from an image is a difficult task as some people might age look quite younger to their respective age or some may look more older to their age. In the proposed system the dataset using the combination of two datasets faces and facial age. The UTK faces contains large number of images consisting of different orientation, lighting also from different age group and ethnicity. Each image is labelled by age, gender and ethnicity. The second dataset is the facial age 2018 dataset contains 9778 images sorted by age groups ranging from 1-110 age.

b) CNN model Description: The input image before fed to the model is resized to 200x200. The convolution layer starts with number of filer equal to 32 in the first layer and increased in step of 32 until it reached 256. The activation of each convolution is ReLU and the number of convolution layer filters is increased by multiplication of 2. After every convolution layer the averagepooling2d is used to down sample the input. The final dense layer with an attribute of 7 is used to match the number of age classes.

IV. EXPERIMENTAL RESULTS

All the tests run on a single machine running with Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz and NVIDIA GeForce GTX 1650 GPU. The performance of each module is listed below.

A. *Emotion Module*

The Emotion module is testes with CK+ dataset. The number of images used for training is 735 and for testing 246 images. The total number of epoch is 25. The Fig 2. shows the Training and Testing accuracy of the model with respect to each epoch.

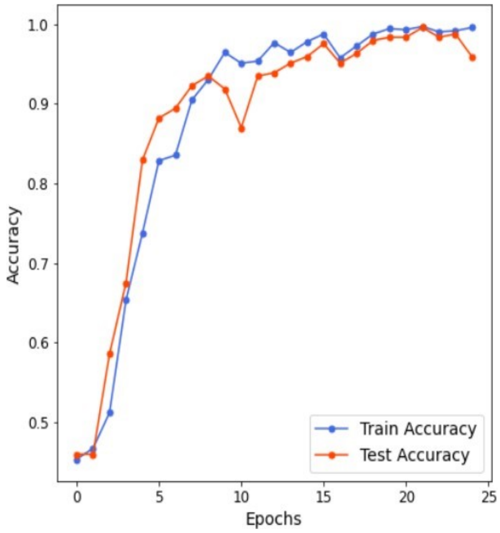


Fig. 2. Accuracy of the Emotion Module

B. *Gender Module*

The Gender detection model is trained on the UTK faces dataset with 17781 training images and 5927 testing im- ages. The Fig 3. shows the accuracy of the module with each epoch.

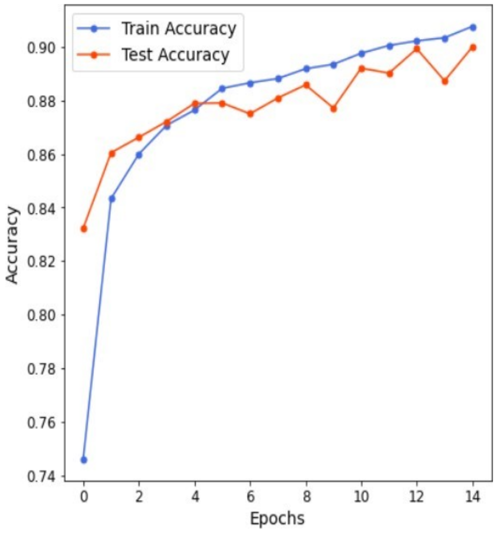


Fig. 3. Accuracy of the Gender Module

C. *Age estimation*

The Age estimation model is trained using the augmented dataset of UTK faces and Facial age dataset. The number of training images is 234400 and the number of images used for testing is 10046.The Fig 4. shows the accuracy of the model with each epoch.

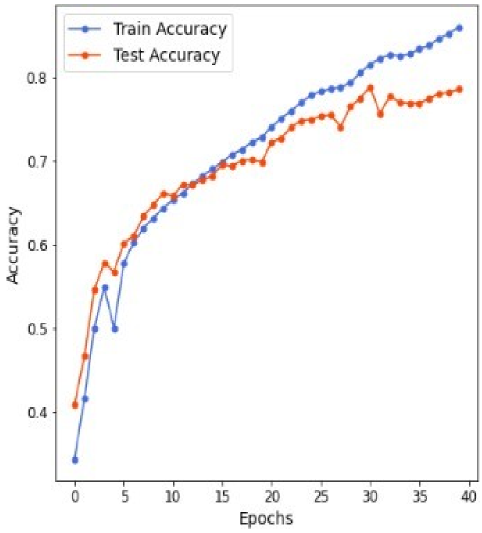
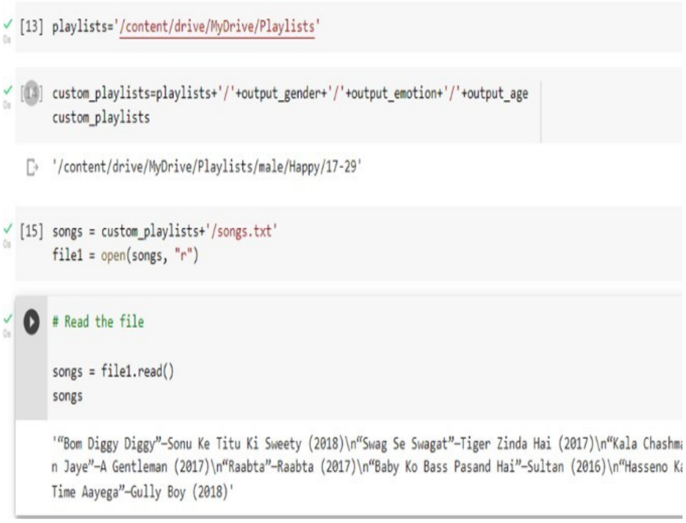


Fig. 4. Accuracy of the Age estimation Module

D. *Music Playlist Recommendation*

The combination of Age, Gender and Emotion detected from the input image is given to find the corresponding playlist. The playlist is custom made just for the illustration. The system seeks the playlists according to the age ,gender and emotion combination in Fig 5.



```
[13] playlists='/content/drive/MyDrive/Playlists'

[14] custom_playlists=playlists+'/'+output_gender+'/'+output_emotion+'/'+output_age
     custom_playlists

     '/content/drive/MyDrive/Playlists/male/Happy/17-29'

[15] songs = custom_playlists+'/songs.txt'
     file1 = open(songs, "r")

     # Read the file

     songs = file1.read()
     songs

     '“Bom Diggy Diggy”–Sonu Ke Titu Ki Sweety (2018)\n“Swag Se Swagat”–Tiger Zinda Hai (2017)\n“Kala Chashma
     n Jaye”–A Gentleman (2017)\n“Raabta”–Raabta (2017)\n“Baby Ko Bass Pasand Hai”–Sultan (2016)\n“Hasseno Ka
     Time Aayega”–Gully Boy (2018)'
```

Fig. 5. Music playlist for the combination

V. Conclusion

The proposed system detects the Age, Gender, and Emotion with the accuracy of 78.9%, 92.1% and 95% respectively. The results from each model is used to find the corresponding playlists. The accuracy of the models can be further improved with larger datasets and the recommended playlists can be integrated with Spotify Music, Gaana or YouTube for wider range of music and real-time playback.

References

[1] X. Xu, Z. Ruan and L. Yang, ”Facial Expression Recognition Based on Graph Neural Network,” 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), 2020, pp. 211-214.

[2] L. Sun, J. Dai and X. Shen, ”Facial emotion recognition based on LDA and Facial Landmark Detection,” 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE), 2021, pp. 64-67.

[3] M. Murtaza, M. Sharif, M. AbdullahYasmin and T. Ahmad, ”Facial expression detection using Six Facial Expressions Hexagon (SFEH) model,” 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0190-0195.

[4] D. Y. Choi, D. H. Kim and B. C. Song, ”Recognizing Fine Facial Micro-Expressions Using Two-Dimensional Landmark Feature,” 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 1962-1966.

[5] S. Roy Supta, M. Rifath Sahriar, M. G. Rashed, D. Das and R. Yasmin, ”An Effective Facial Expression Recognition System,” 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2020, pp. 66-69.

[6] K. Zhang et al., ”Age Group and Gender Estimation in the Wild With Deep RoR Architecture,” in IEEE Access, vol. 5, pp. 22492-22503..

[7] H. Kondo and F. N. Kondo, ”Convolutional Neural Networks on Multichannel Time Series of Smartphone Applications for Gender or Age Range Classification,” 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), 2020, pp. 522-525.

[8] E. P. Ijjina, G. Kanahasabai and A. S. Joshi, ”Deep Learning based approach to detect Customer Age, Gender and Expression in Surveillance Video,” 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6.

[9] S. T. Rahman, A. Arefeen, S. S. Mridul, A. I. Khan and S. Subrina, ”Human Age and Gender Estimation using Facial Image Processing,” 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1001-10059.

[10] P. Smith and C. Chen, ”Transfer Learning with Deep CNNs for Gender Recognition and Age Estimation,” 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2564-2571.

[11] D. Ayata, Y. Yaslan and M. E. Kamasak, ”Emotion Based Music Recommendation System Using Wearable Physiological Sensors,” in IEEE Transactions on Consumer Electronics, vol. 64, no. 2, pp. 196-203, May 2018.

[12] V. P. Sharma, A. S. Gaded, D. Chaudhary, S. Kumar and S. Sharma, ”Emotion-Based Music Recommendation System,” 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021.

[13] S. Gilda, H. Zafar, C. Soni and K. Waghurdekar, ”Smart music player integrating facial emotion recognition and music mood recommendation,” 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 154-158.

[14] R. Chinmayi, N. Sreeja, A. S. Nair, M. K. Jayakumar, R. Gowri and A. Jaiswal, ”Emotion Classification Using Deep Learning,” 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1063-1068.

[15] S. Ramakrishnan, N. Upadhyay, P. Das, R. Achar, S. Palaniswamy and A. A. Nippun Kumaar, ”Emotion Recognition from Facial Expressions using Images with Arbitrary Poses using Siamese Network,” 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 268-273.

[16] P. Supriya, R. Jayabarathi, C. Jeyanth, Y. Ba, A. Sarvesh and M. Shurfudeen, ”Preliminary Investigation for Tamil cine music deployment for mood music recommender system,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 1111-1115.

[17] Devi Babu and Supriya.P P, ”Music Emotion Recommender System using Spectral Features-a Malayalam cine music deployment,” 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1306-1310.