SENTIMENT BASED MUSIC RECOMMENDATION SYSTEM

Submitted by,

ASWATHY LOFY RAJ

(Register Number:233011)

ATHIRA.S

(Register Number:233012)

CHRISTA DAVIS

(Register Number: 233112)

DIKHIL SABU VARGHESE

(Register Number:233113)

GREESHMA.AR

(Register Number:233305)

School of Digital Sciences

Kerala University of Digital Sciences, Innovation and Technology

(Digital University Kerala)

ACKNOWLEDGEMENT

Our heartfelt and sincere thanks go out to Dr. Sanil. P. Nair, Professor at the Digital University Kerala in Trivandrum, who served as our mentor and played a pivotal role in helping me successfully complete our project. I would also like to extend my sincere thanks to the members of my project team who worked tirelessly and collaboratively, demonstrating unwavering dedication and expertise that were instrumental in achieving our project's goals. I deeply appreciate the collective effort and shared commitment that made this project a resounding success.

CONTENT

1.ABSTRACT

Automakers have the challenge of finding new technologies that will naturalize themselves in users' lives while improving the driver experience. This work describes an advanced in-car entertainment system that makes use of video analysis for emotion-based song recommendations. These technologies allow the advanced in-car system to understand, in real-time, the emotional state of the driver.

Our system uses OpenCV for video capture and processing. We focus much on the cabin area of the vehicle, where the driver's face can be captured. In this paper, we are going to work with the best-presented facial detection algorithms so far in the market and very high-tech deep learning models in order for the driver's face to be recognized and analysed very precisely in determining his emotional phase, for instance, happiness and sadness and anger. Such real-time emotion analytics provide a possibility for the development of a personalized and responsive music recommendation system.

The ultimate vision of our project lies in how well this technology could be integrated into vehicles to change the in-car entertainment experience. By knowing the emotions of the driver, this could call for an integration of music that would make a drive much better and free from stress. This would relieve a stressed driver by suggesting some calming music and would stimulate a vibrant driver by suggesting some cheerful music.

Therefore, it diverts the driver less and has much leeway for driver safety, such as the contextually relevant and emotionally tailored recommendations of the kind of music to be played. The total driving experience is loaded with much more adaptability and responsiveness that again adds to both safety and pleasure in driving.

In other words, the song recommendation system was brought in with a mix of computer vision powers along with deep learning and Convolutional Neural Networks, leading to a very sophisticated in-car entertainment solution based on emotions. Our aim is that with such an understanding and hence the reaction in respect to the emotions of the driver, the future definition of car entertainment becomes much safer, highly personalized, and emotionally engaging while driving.

2.INTRODUCTION

In today's fast world, getting the ideal song that matches our mood and current feelings is a task in itself. Our novel project is working toward designing a system that is supposed to pick songs, based on the emotional state of the driver, which is intuitional.

Employing high-end technologies such as OpenCV, Python, deep learning, Convolutional Neural Networks, face detection, and emotion recognition, among others, we would want to improve the music experience in vehicles. The system helps to track the emotions of the driver and, in real time, it suggests music which correlates with the current state of emotions of the driver, thus making driving not only safe but more pleasant.

We capture and process the video feeds from an in-car camera that focuses on the driver's face via OpenCV. This will be fed to a neural network that will be trained for face detection and extraction of the facial features. In our multitask learning approach, this CNN is trained on extensive datasets of annotated facial images displaying different emotions to recognize the face of the driver and decipher his emotions.

Accurate emotion recognition can make relevant music recommendations in varied contexts; for example, suggesting joyful songs when joy is detected, or playing soothing music in instances of stress or sadness. This technology personalizes the driving experience according to tastes and preferences, while safety priorities are maintained by limiting distraction.

The project goes far beyond personal cars; it can also enhance ride-sharing services, autonomous vehicles, and public transportation, making entertainment mood-driven in a new way. Our project will not only change the way we experience music on the road but also make it much more personal and responsive to our emotional needs.

3. LITERATURE REVIEW

Interest in emotion-based music recommendation systems has grown over the past years because such a system surely can enhance the user's experience by matching musical selections with the emotional state of the listeners. Earlier research in this area was mostly oriented towards manual approaches for emotion detection and music categorization. Current advances in machine learning, however, permit more sophisticated methods with large impacts in the domains of computer vision and audio processing.

Emotion Detection Techniques: Traditional methods were based on explicit inputs by the users or other simple physiological signals. The more covert methods, like facial expression analysis using algorithms of the Haar Cascade Classifier, have been the subject of recent research. Viola and Jones developed in 2001 an algorithm of emotion detection that has been applied in music recommendation systems. Systems can recognize emotions like as sadness, happiness, anger, surprise, and fear by training classifiers on facial expression datasets, and then select music to match these emotional states.

Audio Feature Extraction: It is important, however, to extract relevant data from the music to make the recommendation process more accurate. The Mel-Frequency Cepstral Coefficients and spectral features, in general, and rhythmic patterns are frequently used in emotion-based recognition of music. These features aid in developing a profile of music from these features that can then be measured against the presence of different identified emotional states of the users.

Machine Learning Models: To improve the accuracy of emotion-based suggestions, a variety of machine learning models were used. SVMs and CNNs have been especially useful. For example, Sriraj Katkuri (2023) suggested a system that uses photos to capture the user's emotional state, employing CNNs for feature extraction and SVMs for emotion recognition. In comparison to traditional methodologies, their approach demonstrated considerable gains in suggestion accuracy and user satisfaction.

Evaluation Metrics: Accuracy, recall, and F1-score are some common metrics that would show the efficacy of an algorithm for music recommendation based on emotion. These variables contribute to determining how well a system can match music tracks according to users' emotional states. Studies have shown that their development using user feedback makes them more responsive to the taste of every individual and increases overall accuracy.

4. PROPOSED METHODOLOGY

Wider Face Dataset

The Wider dataset consists of 32,203 images and 393,703 face The wider bounding boxes with a high degree of variability in scale, pose, expression, occlusion and illumination. The wider face dataset is split into training(40%), validation(10%) and testing (50%) subsets by randomly sampling from 61 scene categories Based on the detection rate of EdgeBox, three levels of difficulty (i.e. Easy, Medium and Hard) are defined by incrementally incorporating hard samples. We define five levels of face image quality and annotate five facial landmarks (i.e. eye centres, nose tip and mouth corners) on faces that can be annotated from the wider face training and validation subsets in total we have annotated 84.6k faces on the training set and 18.5k faces on the validation set.

Face Dataset

AffectNet Dataset

Facial affect database is created from the internet by querying different search engines using 1250 emotion related tags in six different languages(English, Spanish, Portuguese, German, Arabic, and Farsi). AffectNet contains more than one million images with faces and extracted facial landmark points. Twelve human experts manually annotated 450,000 of these images in both categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face. Fig 2 shows sample images from AffectNet and their valence and arousal annotations. To calculate the agreement level between the human labelers, 36,000

images were annotated by two human labelers. AffectNet is by far the largest database of facial affect in still images which covers both categorical and dimensional models. The cropped region of the facial images, the facial landmark points, and the affect labels are publicly available to the research community.

Fig2: Samples of queried images from the web and their annotated tags

Face Detection using Retina Face

We use retina face detection trained on wider face dataset for detecting face The "Wider face" dataset is a popular benchmark dataset used for training and evaluating face detection algorithms. It is widely used in the computer vision community for developing and testing face detection models.

The RetinaFace architecture is chosen as the core model for face detection in this project. RetinaFace is renowned for its ability to detect faces at multiple scales and orientations, making it suitable for handling various real-world scenarios. The architecture comprises of a pre trained convolutional neural network (CNN), such as ResNet or MobileNet, serves as the backbone for feature extraction. This network is responsible for capturing discriminative features from input images. A Feature Pyramid Network (FPN) is incorporated to create a feature pyramid that combines features from different levels of the backbone network. This enables the model to detect faces at various scales. The model utilizes anchor boxes (prior boxes) with different aspect ratios and scales, placed at multiple positions in the feature pyramid. These anchor boxes act as reference boxes for both classification and regression tasks. One part of the model focuses on binary classification, determining whether anchor boxes contain faces or not. It assigns confidence scores to each anchor box,

indicating the likelihood of it containing a face. Bounding Box Regression Subnet is responsible for regressing the coordinates of bounding boxes around detected faces. It refines anchor boxes to closely align with actual faces.

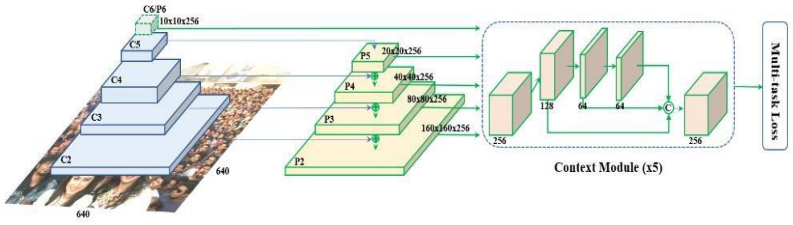

Fig3: Retina Face

Emotion recognition using multi task network

In this project we use a multi task network which is a lightweight Cnn for detecting emotions in a face. This multi network is fine tuned using many datasets to solve several facial attributes recognition problems. The disjoint features among the tasks are exploited to increase the accuracies. At first the base CNN is pre-trained on face identification using very large VGGFace2 dataset.

We use architectures such as MobileNet, EfficientNet and RexNet as a backbone face recognition network. The resulted neural net extracts facial features that are suitable to discriminate one subject from another. These features can be used to predict the attributes that are stable for a given person. The CNN is further fine-tuned on emotion dataset to use valuable information about facial features in order to predict the facial attributes that are orthogonal to the identity.

The CNNs are trained sequentially starting from face identification problem and further tuning on different facial attribute recognition tasks. At first, the face recognition CNN is trained using the VGGFace2 dataset. The training set contain 3,067,564 photos of 9131 subjects, while the remaining 243,722 images fill the testing set. The new head, i.e., FC layer with 9131 outputs and softmax activation, was added to the network pre-trained on ImageNet. The weights of the base net were frozen and the head was learned during 1 epoch. The categorical cross-entropy loss function was optimized using contemporary SAM (Sharpness-Aware Minimization) and Adam with learning rate equal to 0.001. Next, the whole CNN is trained in 10 epochs in the same way but with learning rate 0.0001. Next, separate heads for age, gender and ethnicity prediction were added and their weights were learned. The training dataset was populated by 300K frontal cropped facial images from the IMDB-Wiki dataset to predict age and gender. Finally, the network is fine-tuned for emotion recognition on the AffectNet dataset. The training set provided by the

authors of this dataset contains 287,651 and 283,901 images for Ce = 8 classes (Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt) and 7 primary expressions (the same without Contempt), respectively. The official validation set consists of 500 images per each class, i.e. 4000 and 3500 images for 8 and 7 classes. We rotate the facial images to align them based on the position of the eyes but without data augmentation. There are two ways to classify 7 emotions were namely, train the model on reduced training set with 7 classes or train the model on the whole training set with 8 classes, but use only 7 scores from the last (Softmax) layer. In both cases, the weighted categorical cross-entropy (softmax) loss was optimized. We use this pretrained model to predict emotions in our project.

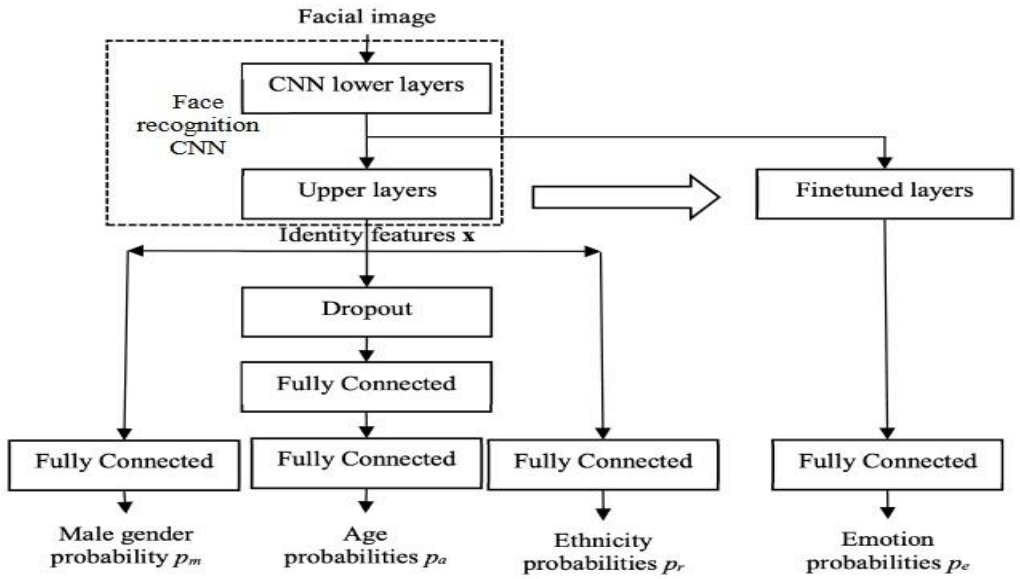

Fig4:Multi Task Network

Song recommendation using data_moods.csv

Data_moods is a dataset containing data about music. We use this data to recommend songs based on the mood/emotion of the user. It has about 686 rows and 19 columns. We recommend this songs from this dataset by querying. We query it so that the songs corresponding to certain moods are recommended to the user in random. This is possible because of the existence of the column mood in this dataset. This dataset contain attributes of the music. These are the features of the dataset :

- name-name of the song
- album-name of the album.
- artist-The name of the artist.
- id- The spotify id for the track.
- release_date-The date the song has been released.
- popularity- The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.
- length- The duration of the track in milliseconds.
- danceability- Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat

strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

- acousticness- A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- energy-Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
- instrumentalness-Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- liveness-Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- valence-A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- loudness-The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
- speechiness-Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other nonspeech-like tracks.
- tempo-The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- key- The key values correspond to the 12 different musical keys in Western music. It’s values ranges from 1-11.
- mood-This attribute contain the mood of the song which are sad, happy, energetic and calm.

5. CONCLUSION

Our project, Sentiment-Based Music Recommendation through Video, is a combination of technology, emotion, and entertainment that can really change the future of music experience in our new digital world. Music is a universal language of emotion, and it has grown and evolved significantly. We are on the brink of creating a future where our vehicles become intuitive companions, curating soundtracks that resonate with our innermost feelings.

Ours is an epitome of what comes to the fore when human emotions and technology are brought together. With the use of the Convolutional Neural Network in training this understanding of feelings and with facial expressions to joy, sadness, anger, and surprise, we could develop a music recommendation system appropriate to each driver's context to enhance his driving experience while ensuring safety for each trip.

We employ a strong single-stage face detector, RetinaFace, of which the performance achieves 1.1% better average precision (AP) over the state-of-the-art with an AP of 91.4%. At first, RetinaFace realizes pixel-wise localization of a face at different scales, and then it detects a person's face and passes it to the entrance of the multitask network with the MobileNet backbone for face recognition. This fine-tuned network demonstrates near state-of-the-art results in predicting emotion. We recommend tracks based on the mood detected with the dataset data_moods.

This is in addition to the obvious excitement of deploying the technology within vehicles. This is where celebration demands peppy numbers and introspection requires soothing melodies, in a perfect union between technology and emotion. Your car will be as much in tune with your state of mind as you are with your environment, making sure that every mood is provided with its perfect musical journey. But then safety is still uppermost, and the system has been designed to work without much distraction, keeping the driver's eye mostly on the road.

The emotion-based song recommendation system using video showcases a few of the infinite possibilities emerging from the confluence of technology, deep learning, and human emotions. Our vision is to revolutionize music-on-the-road experiences with personalized, responsive, and meaningful music experiences.

6. FUTURE ENHANCEMENT

Emotion-based music recommendation systems would someday fundamentally change the way people discover and relate to music. Several avenues for further improvements open up with the advancement in technology. We can also upgrade our video-based music recommendation system to not only suggest music to a user but play it automatically. Such music recommendation systems can be even more tailored if they consider the user's present emotional state and his/her past emotional patterns. It is even possible to track the emotional changes a user exhibits towards songs over a long period and then adapt the recommendations based on the change. For instance, future systems could utilize cross-cultural and multilingual emotional analysis for culturally aware recommendations. In addition, context of the user such as location, activities, and social relations can be used to enhance emotion-based recommendations.

7. REFERENCES

1.Brijesh Bakariya,Arshdeep Singh,Pankaj Raju,Rohit Rajpoot,Krishna kumar Mohbey(2024),Facial Emotion recognition and music recomendation system using CNN based deep learning technique https://doi.org/10.1007/s12530-023-09506-z

2. Eva sarin,Srishti Vashishtha,Megha,Simran Kaur (2021),4th international conference on recent trends in computer science and technology (ICRTCST), SentiSpotMusic : A music recommendation system based on Sentiment analysis 10.1109/ICRTCST54752.2022.9781862

3. Benamara NK et al (2021) Real-time facial expression recognition using smoothed deep neural network ensemble.integrat Comput-Aided https://doi.org/10.3233/ICA-200643

4. V Tejaswini Priyanka, Y Reshma Reddy,G Ramesh,S Gomathy (2023),7th International Conference on intelligent Computing and control system(ICICCS),A Novel Emotion based Music Recommendation System Using CNN 10.1109/ICICCS56967.2023.10142330