EMOTION BASED MUSIC RECOMMENDATION SYSTEM

Submitted by,

HIBA HYDER

(Register Number: 223030)

JOEL CHRISTY FRANCIS

(Register Number: 223031)

FATHIMA SHABEERA

(Register Number: 223024)

HARSHIN RONALDO

(Register Number: 223029)

School of Digital Sciences

Kerala University of Digital Sciences, Innovation and Technology

(Digital University Kerala)

ACKNOWLEDGEMENT

My heartfelt and sincere thanks go out to Dr. T.K. Manoj Kumar, Professor at the Digital University Kerala in Trivandrum, who served as my 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

As technology continues to integrate with our daily lives, the automotive industry seeks innovative ways to enhance the driving experience. In this context, we present a novel approach to in-car entertainment through Emotion-Based Song Recommendation using video analysis. Leveraging computer vision, deep learning, and Python programming, we aim to understand and respond to the emotional state of the driver in real-time.

Our system utilizes OpenCV for video capture and processing, enabling the capture of the driver's face within the car's cabin. Employing state-of-the-art face detection deep learning algorithms, we accurately identify the driver's face. Then emotion recognition is achieved through advanced deep learning model. By analyzing facial crop expressions, we can determine the emotional state of the driver, including emotions such as happiness, sadness, anger, and more. This real-time emotion analysis is essential for delivering a personalized and responsive music recommendation system.

The end goal of our project is to seamlessly integrate this technology into vehicles, transforming the in-car entertainment experience. By understanding the driver's emotions, we can suggest music that aligns with their mood and preferences, creating a more enjoyable and stress-free driving experience. For instance, a driver displaying signs of stress may receive calming music recommendations, while a cheerful driver may be offered upbeat tunes to enhance their journey.

Our system's adaptability and responsiveness ensure that it prioritizes driver safety, keeping distractions to a minimum. In addition, it offers a personalized and emotionally tailored entertainment experience, enhancing the overall driving experience.

In summary, our Emotion-Based Song Recommendation system leverages the power of computer vision ,deep learning and CNN to create a cutting-edge in-car entertainment solution. By understanding and responding to the driver's emotions, we aim to provide a safer, more enjoyable, and personalized driving experience, ultimately redefining the future of automotive entertainment.

2. INTRODUCTION

In today's fast-paced world, technology continues to shape our daily lives in countless ways, and one of the most profound impacts can be seen in the realm of entertainment. Music, in particular, has undergone a remarkable transformation in the digital age. With the vast amount of music available at our fingertips, the challenge now lies in finding the perfect song that resonates with our current mood and emotions. Imagine a world where your car becomes your personal DJ, intuitively selecting songs that match your emotional state. This is the future we are striving to achieve with our innovative project - Emotion-Based Song Recommendation using video.

In this digital era, we are leveraging the power of cutting-edge technologies, including OpenCV, Python, deep learning, Convolutional Neural Networks (CNNs), face detection, and emotion recognition, to create a unique and immersive music experience within the confines of your vehicle. Our project's ultimate goal is to deploy a robust system that analyzes the emotions of the driver in real-time and suggests music that complements their current emotional state, ensuring a safer and more enjoyable driving experience.

At the core of our project lies the fusion of computer vision and deep learning techniques. We employ OpenCV, an open-source computer vision library, to capture and process video feeds from an in-car camera, which is strategically positioned to focus on the driver's face. This video feed is then fed into a complex neural network architecture that excels at detecting faces and extracting facial features with remarkable precision.

The heart of our system is our multi-task learning approach, which not only identifies the driver's face but also deciphers the emotions etched upon it. Through deep learning, our Convolutional Neural Network has been trained to understand human emotions by examining facial expressions, including joy, sadness, anger, surprise, and more. This is achieved by training the network on extensive datasets containing thousands of annotated facial images displaying various emotional states.

The ability to accurately recognize the driver's emotions is crucial in ensuring that the music recommendations are not only tailored but also contextually relevant. For instance, a cheerful and upbeat song may be recommended if the system detects joy on the driver's face, while soothing and mellow tunes might be selected if it detects signs of stress or sadness.

The deployment of this technology in a vehicle brings an exciting dimension to our project. Imagine embarking on a long road trip, and as you navigate the highways and byways, your car senses your emotional fluctuations and responds accordingly. During moments of celebration, it cranks up the tempo with lively tunes. In times of

contemplation, it soothes your soul with melodic ballads. The car essentially becomes an extension of your emotional state, curating a musical journey that complements your feelings.

Safety is, of course, a paramount concern when implementing such advanced technologies in a moving vehicle. Rest assured, our system has been designed with safety as a top priority. The video analysis is performed in real-time, and the music recommendations are generated with minimal distractions, ensuring that the driver's focus remains primarily on the road.

The potential applications of emotion-based song recommendation extend far beyond the realm of personal vehicles. This technology can also be integrated into ride-sharing services, autonomous vehicles, and public transportation systems, enhancing the overall passenger experience and providing an innovative approach to mood-driven entertainment.

In conclusion, our journey into the world of emotion-based song recommendation using video is a testament to the limitless possibilities that arise from the convergence of computer vision, deep learning, and human emotions. Our vision is to transform the way we experience music on the road, making it more personalized and responsive to our emotional needs. As we move forward, we are excited to share our progress and discoveries on this exciting path towards a future where our cars truly understand and cater to our emotions through the power of music.

3. LITERATURE REVIEW

This category reviews completed multiple studies and theses in the field.

Many academics are working to discover a method where our car becomes our personal DJ, intuitively selecting songs that match our emotional state. With this goal we made the emotion based music recommendation system. For this we used the cutting-edge technologies like OpenCV, Python, and Convolutional Neural Networks (CNNs) to create an immersive in-car music experience. Our goal is to develop a system that analyzes the driver's real-time emotions and recommends music that suits their mood, enhancing both safety and enjoyment.Our system integrates computer vision and deep learning . OpenCV captures and processes video from an in-car camera focused on the driver's face. A sophisticated neural network detects faces and deciphers emotions by studying facial expressions. Deep learning trains the Convolutional Neural Network on extensive datasets with various emotional states. Accurate emotion recognition ensures contextually relevant music recommendations. For instance, joyful music plays when joy is detected, while calming tunes are suggested during moments of stress or sadness.This technology transforms road trips, adapting music to your emotions as you drive. Safety is paramount; video analysis and music recommendations minimize distractions. Drivers can override the system if needed.Emotion-based song recommendation isn't limited to personal vehicles. It can enhance ride-sharing, autonomous vehicles, and public transportation, providing mood-driven entertainment.

For the face detection we used the RetinaFace model and MobileNet deep learning architecture. A work based on RetinaFace was published on 4 May 2019 by Jiankang Deng, Jia Guo , Middlesex University London, under the title-‘RetinaFace: Single-stage Dense Face Localisation in the Wild’[**1**].This paper present the robust single-stage face detector, RetinaFace, which performs pixel-wise face localisation on various scales of faces by taking advantages of joint extra-supervised and self-supervised multi-task learning. They make contributions mainly in five aspects: 1) Manually annotate five facial landmarks on the WIDER FACE dataset and observe significant improvement in hard face detection with the assistance of this extra supervision signal.2) Then add a selfsupervised mesh decoder branch for predicting a pixel-wise 3D shape face information in parallel with the existing supervised branches.3) On the WIDER FACE hard test set, RetinaFace outperforms the state of the art average precision (AP) by 1.1% (achieving AP equal to 91.4%).4) On the IJB-C test set, RetinaFace enables state of the art methods (ArcFace) to improve their results in face verification (TAR=89.59% for FAR=1e-6). 5) By employing light-weight backbone networks, RetinaFace can run real-time on a single CPU core for a VGA-resolution image. The WIDER FACE dataset is a widely used dataset for training and evaluating face detection algorithms. It contains diverse images with

annotated bounding boxes around faces, making it suitable for testing algorithms in real-world scenarios with various lighting conditions and challenges like occlusions and extreme poses. It serves as a benchmark for assessing the accuracy and robustness of face detection models.A research paper about WIDER FACE dataset published on 20 Nov 2015 by Shuo Yang, Ping Luo, Chen Change Loy and Xiaoou Tang from Department of Information Engineering, The Chinese University of Hong Kong, under the title –"WIDER FACE: A Face Detection Benchmark"[**2**].In this they are saying that they introduced the WIDER FACE dataset to facilitate the future face detection research. This dataset is 10 times larger than the preoccured datasets. The dataset contains rich annotations, including occlusions, poses, event categories, and face bounding boxes. Faces in the proposed dataset are extremely challenging due to large variations in scale, pose and occlusion.They benchmark several representative detection systems, providing an overview of state-of-the-art performance and propose a solution to deal with large scale variation. Emotion recognition from face detection is crucial for an emotion-based music recommendation system using video, as it enables personalized and context-aware music suggestions. By analyzing facial expressions, the system can adapt playlists to match the viewer's emotional state, enhancing the overall viewing experience and emotional connection with the content.In a research paper titled-'Facial expression and attributes recognition based on multi-task learning of lightweight neural networks'[**3**],published by Andrey V Savchenko, HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia on 4 oct,2021,it is saying that the multi-task learning of lightweight convolutional neural networks is studied for face identification and classification of facial attributes trained on cropped faces without margins.The necessity for fine-tune these neural-network is to predict facial expressions.Through the experiments it had shown that the usage of the trained models as feature extractors of facial regions in video frames leads to 4.5% higher accuracy than the previously known stateof-the-art single models for the AFEW and the VGAF datasets from the EmotiW challenges.

Our journey into emotion-based song recommendation through video showcases the power of computer vision, deep learning, and human emotions. We aim to revolutionize in-car music, making it more personalized and responsive. We look forward to sharing our progress in a future where cars understand and cater to our emotions through music.

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.

Fig1:Wider 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.

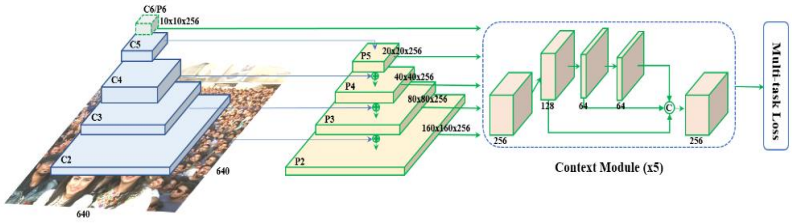

Fig3: Retina Face

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.

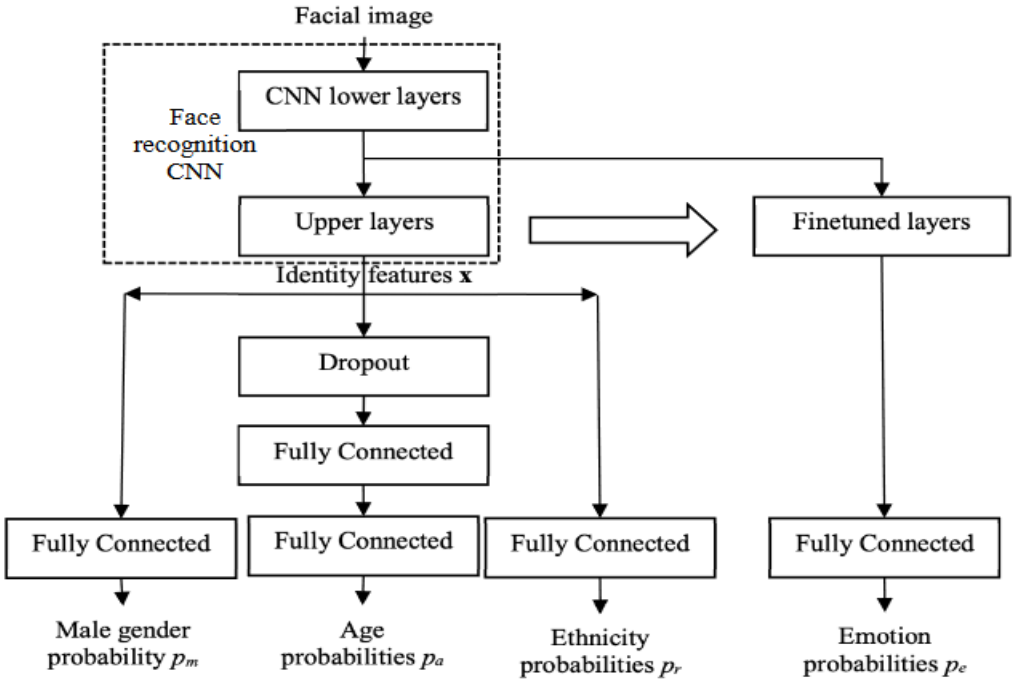

Fig4:Multi Task Network

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 non-speech-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, Emotion-Based Song Recommendation using video, represents a remarkable fusion of technology, emotion, and entertainment that has the potential to redefine how we experience music in the digital age. Music, as a universal language of emotion, has undergone a significant transformation, and we are on the cusp of ushering in a future where our vehicles become intuitive companions, curating soundtracks that resonate with our innermost feelings.

Our system is a testament to the possibilities that arise from the synergy between human emotion and technology. We've trained our Convolutional Neural Network to understand human emotions by scrutinizing facial expressions, encompassing a range of feelings from joy to sadness, anger to surprise. The result is a personalized and contextually relevant music recommendation system, enriching the driving experience and ensuring both enjoyment and safety on the road. We use a robust single-stage face detector, named RetinaFace, RetinaFace outperforms the state of the art average precision (AP) by 1.1% (achieving AP equal to 91.4%) RetinaFace performs pixel-wise face localisation on various scales of faces, and detects the face of the person it is then fed into a multi task network with mobilenet as the backbone face recognition network, this network is trained and finetuned to predict emotions, this network gives near state-of-the-art results and then with the help of a dataset data_moods we recommend song based on the mood of the person.

The deployment of this technology within vehicles opens up an exciting dimension in our project. It is a harmonious fusion of technology and emotion, where celebration calls for lively tunes, and introspection invites soothing melodies. Your car becomes an extension of your emotional state, curating a musical journey that complements your every mood. Safety remains our paramount concern. Our system has been meticulously designed to operate seamlessly, with minimal distractions, ensuring that the driver's focus remains firmly on the road.

The emotion-based song recommendation system using video is a testament to the limitless possibilities that emerge at the intersection of technology, deep learning, and human emotions. Our vision is to revolutionize the way we experience music on the road, making it more personalized, responsive, and deeply meaningful. As we continue forward on this exciting path, we are eager to share our progress and discoveries, moving toward a future where our vehicles truly understand and cater to our emotions through the enchanting language of music.

6. FUTURE ENHANCEMENT

Emotion-based music recommendation systems have the potential to revolutionize the way people discover and interact with music. As technology continues to advance, we can have many improvements in certain areas. In our video based music recommendation system we can improve the system so that it not only recommends but also play the music. Emotion-based recommendation systems can become even more personalized by considering not only the user's current emotional state but also their historical emotional preferences. This could involve analyzing a user's emotional response to songs over time and adjusting recommendations accordingly. Emotions are expressed differently across cultures and languages. Future systems may incorporate cross-cultural and close-lingual emotion analysis to provide recommendations that are sensitive to cultural nuances. Emotion-based recommendations can benefit from considering the user's context, such as location, activity, and social interactions.

7. REFERENCES

1. Jiankang Deng, Jia Guo, Jinke Yu, Imperial College London, Middlesex University London(2019). RetinaFace: Single-stage Dense Face Localisation in the Wild. https://arxiv.org/pdf/1905.00641v2.pdf
2. Shuo Yang, Ping Luo, Chen Change Loy, Xiaoou Tang, Department of Information Engineering, The Chinese University of Hong Kong(2015). WIDER FACE: A Face Detection Benchmark. https://arxiv.org/pdf/1511.06523.pdf
3. Ali Mollahosseini, Student Member, IEEE, Behzad Hasani, Student Member, IEEE, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. http://mohammadmahoor.com/wp-content/uploads/2017/08/AffectNet_oneColumn-2.pdf
4. Andrey V. Savchenko HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia(2021). Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. https://arxiv.org/pdf/2103.17107.pdf