Recommendation of Music Based on Facial Emotion using Machine Learning Technique

[1]G Sakthi Priya, [2]A Evangelin Blessy, [3]S Jeya Aravinth, [4]M Vignesh Prabhu and [5]R VijayaSarathy
[1,2,3,4,5] Department of CSE, Ramco Institute of Technology, Rajapalayam, TamilNadu, India.
[1] gsakthimaheswari@gmail.com ,[2] evebless8@gmail.com, [3] 953621104019@ritrjpm.ac.in, [4]953621104058@ritrjpm.ac.in
[5]953621104059@ritrjpm.ac.in

Article Info
A. Haldorai et al. (eds.), *2nd International Conference on Materials Science and Sustainable Manufacturing Technology,* Advances in Computational Intelligence in Materials Science.
Doi: https://doi.org/10.53759/acims/978-9914-9946-9-8_16

Abstract—Music plays a vital role in human life, and it is a valid therapy to potentially reduce depression, anxiety, as well as to improve mood, self-esteem, and quality of life. Music has the power to change human emotion as expressed through facial expression. It's a difficult task to recommend music based on emotion. The existing system on emotion recognition and music recommendation is focused on depression and mental health analysis. Hence a model is proposed to recommend music based on recognition of face expression to improve or change the emotion. Face emotion recognition (FER) is implemented using YoloV5 algorithm. The output of FER is a type of emotion classified as happy, anger, sad, and neutral which is the input to music recommendation system. A Music player is created to keep track of the user's favorite based on the emotion. If the user is new to the system, then generalized music will be suggested. The aim of the paper is to recommend music to the user according to their emotion to further improve it.

Keywords—Face Emotion Recognition, Music Recommendation System, Yolov5, Machine Learning Technique, Emotion Classification, Songs.

I. INTRODUCTION

Music has the ability to evoke powerful emotional responses such as chills and thrills in listeners. Positive emotions dominate musical experiences. Pleasurable music may lead to the release of neurotransmitters associated with reward, such as dopamine. Listening to music is an easy way to alter mood or relieve stress [1]. People use music in their everyday lives to regulate, enhance, and diminish undesirable emotional states (e.g., stress, fatigue).

The enjoyment of music appears to involve the same pleasure center in the brain as other forms of pleasure, such as food and drugs. Evidence shows that an aesthetic stimulus, such as music, can naturally target the dopamine systems of the brain that are typically involved in highly reinforcing and addictive behaviors.

Music can be experienced as pleasurable both when it fulfills and violates expectations [2]. The more unexpected the events in music, the more surprising is the musical experience which appreciate music that is less predictable and slightly more complex.

Music doesn't only evoke emotions at the individual level, but also at the interpersonal and intergroup level. Listeners mirror their reactions to what the music expresses, such as sadness from sad music, or cheer from happy music. Similarly, ambient music affects shoppers' and diners' moods [3].

Emotion based music player is a novel approach that helps the user to automatically play songs according to the emotions of the user. It recognizes the facial emotions of the user and plays the songs according to their emotion. The emotions are recognized using a machine learning method YoloV5 model. This project is fully focused on training a YoloV5 model for emotion detection and improves the accuracy of face emotion detection.

II. RELATED WORK

A fast and accurate system for face detection, identification, and verfication

The availability of large annotated datasets and affordable computation power, according to [4] has led to impressive improvements in the performance of convolutional neural networks (CNNs) on various face analysis tasks [5]. They described a deep learning pipeline for unconstrained face identification and verification that achieves state-of-the-art performance on a variety of benchmark datasets in this paper. They detailed the design of the various modules involved in automatic face

recognition, including face detection, landmark localization and alignment, and face identification/verification. Deep pyramid single shot face detector (DPSSD), which is fast and detects faces with large scale variations, was proposed (especially tiny faces). They also proposed a new loss function called crystal loss for face verification and identification tasks. Crystal loss confines the feature descriptors to a fixed-radius hypersphere, minimizing the angular distance between positive subject pairs while increasing the angular distance between negative subject pairs. They presented results of the proposed face detector evaluation on challenging unconstrained face detection datasets. The researchers then presented experimental results for end-to-end face verification and identification on IARPA Janus Benchmarks A, B, and C (IJB-A, IJB-B, and IJB-C), as well as the Janus Challenge Set 5. (CS5).

Music recommendation system based on user's sentiments extracted from social networks

In [6] proposed sentiment analysis, which has been investigated by several Internet services to recommend contents based on human emotions expressed through informal texts posted on social networks [7]. However, the sentiment analysis metrics only classify a sentence as positive, neutral, or negative in intensity and do not detect sentiment variations based on the user's profile. This paper presents a music recommendation system based on a sentiment intensity metric, dubbed enhanced Sentiment Metric (eSM), which combines a lexicon-based sentiment metric with a correction factor based on the user's profile. Subjective tests in a laboratory setting are used to identify this correction factor. The correction factor is developed based on the experimental results and is used to adjust the final sentiment intensity. Users' sentiments are extracted from sentences posted on social networks, and the music recommendation system is carried out using a low-complexity framework for mobile devices, which suggests songs based on the current user's sentiment intensity. Furthermore, the framework was designed with usability ergonomics in mind. The proposed framework's performance is evaluated with remote users using the crowdsourcing method, yielding a rating of 91 percent user satisfaction, outperforming a randomly assigned song suggestion, which yielded a rating of 65 percent user satisfaction. Furthermore, the paper presents low perceived impacts on the analysis of energy consumption, network, and latency in accordance with the recommendation system's processing and memory perception, demonstrating benefits for the consumer electronic world.

Yolo only looked once: unified, real-time object detection.

The system [8] was proposed using a novel object detection approach [9]. Previous work on object detection repurposes classifiers to detect objects. They framed object detection instead as a regression problem to spatially separated bounding boxes and associated class probabilities. In a single evaluation, a single neural network predicts bounding boxes and class probabilities directly from full images. Because the entire detection pipeline is a single network, detection performance can be optimized end-to-end. Their unified architecture is lightning-fast. Their YOLO model, at 45 frames per second, processes images in real time. Fast YOLO, a smaller version of the network, processes an astounding 155 frames per second while achieving twice the mAP of other real-time detectors. YOLO makes more localization errors than state-of-the-art detection systems but is less likely to predict false positives on background. YOLO learns very general object representations. When generalizing from natural images to other domains such as artwork, it outperforms other detection methods such as DPM (Deformable Parts Model) and R-CNN.

YOLOv4: optimal speed and accuracy of object detection

In [10] proposed a large number of features to improve the accuracy of Convolutional Neural Networks (CNNs) [10]. Practical testing of such feature combinations on large datasets is required, as is theoretical justification of the results. Some features are only applicable to certain models and problems, or only to small-scale datasets; whereas others, such as batch-normalization and residual-connections, are applicable to the vast majority of models, tasks, and datasets. Weighted-Residual-Connections (WRC), Cross-Stage-Partial-Connections (CSP), Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT), and Mish-activation are assumed to be such universal features. They combined new features such as WRC, CSP, CmBN, SAT, Mish activation, Mosaic data augmentation, CmBN, DropBlock regularization, and CIoU loss to achieve state-of-the-art results: 43.5 percent AP (65.7 percent AP50) for the MS COCO dataset at a real-time speed of 65 FPS on Tesla V100.

SSD: single shot multibox detector

In [11] proposed using a single deep neural network to detect objects in images. SSD discretizes the output space of bounding boxes into a set of default boxes with different aspect ratios and scales per feature map location. At prediction time, the network computes score for the presence of each object category in each default box and adjusts the box to better match the object shape. Furthermore, the network naturally handles objects of varying sizes by combining predictions from multiple feature maps with different resolutions. SSD is simpler than object proposal methods because it eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. This makes SSD simple to train and integrate into systems that require a detection component. SSD has competitive accuracy to methods that use an additional object proposal step and is much faster, according to experimental results on the PASCAL VOC, COCO, and

ILSVRC datasets. SSD also provides a unified framework for both training and inference. On the VOC2007 test at 59 FPS on an Nvidia Titan X, SSD achieves 74.3 percent mAP1 for 300 input and 76.9 percent mAP for 512 input, outperforming a comparable state-of-the-art Faster R-CNN model. SSD has much better accuracy than other single stage methods, even with a smaller input image size.

YOLO9000: better, faster, stronger

YOLO9000, [11] proposed is a cutting-edge, real-time object detection system capable of detecting over 9000 object categories. First, they propose several novel and prior work-based improvements to the YOLO detection method. The improved model, YOLOv2, performs admirably on standard detection tasks such as PASCAL VOC and COCO. The same YOLOv2 model can run at varying sizes thanks to a novel, multi-scale training method, providing an easy tradeoff between speed and accuracy. YOLOv2 receives 76.8 mAP on VOC 2007 at 67 FPS. YOLOv2 achieves 78.6 mAP at 40 FPS, outperforming state-of-the-art methods such as Faster RCNN with ResNet and SSD while still running significantly faster. They proposed a method for training on object detection and classification simultaneously. Using this method, they trained YOLO9000 on both the COCO detection and ImageNet classification datasets at the same time. Their collaborative training enables YOLO9000 to predict detections for object classes that lack labelled detection data. On the ImageNet detection task, they validate their approach. Despite only having detection data for 44 of the 200 classes, the YOLO9000 achieves 19.7 mAP on the ImageNet detection validation set. YOLO9000 receives 16.0 mAP on the 156 classes that are not in COCO. In real-time, YOLO9000 predicts detections for over 9000 different object categories.

Bam: Bottleneck Attention Module

The [12] proposed recent advances in deep neural networks developed through architecture search for higher representational power. They focused on the effect of attention in general deep neural networks in this paper. They proposed the Bottleneck Attention Module (BAM), a simple and effective attention module that can be integrated with any feed-forward convolutional neural network. Their module derives an attention map via two distinct channels: channel and spatial. They placed their module at each bottleneck of models where feature map down sampling occurs. Their module builds a hierarchical attention at bottlenecks with a variety of parameters and is trainable end-to-end with any feed-forward model. They validated their BAM using the CIFAR-100, ImageNet-1K, VOC 2007, and MS COCO benchmarks. Their experiments show consistent improvement in classification and detection performance with different models, demonstrating BAM's broad applicability.

III.PROPOSED WORK

Fig 1 depicts the system architecture of face emotion recognition and the music recommendation system. Face Image dataset is given as input to the system. Each face in the images will be marked with boundary line. YoloV5 algorithm is implemented on the preprocessed image dataset which produce the emotion as a result. The detected emotion is given as an input to the MusiCart (a website with musics) to recommend a music according to the emotion.

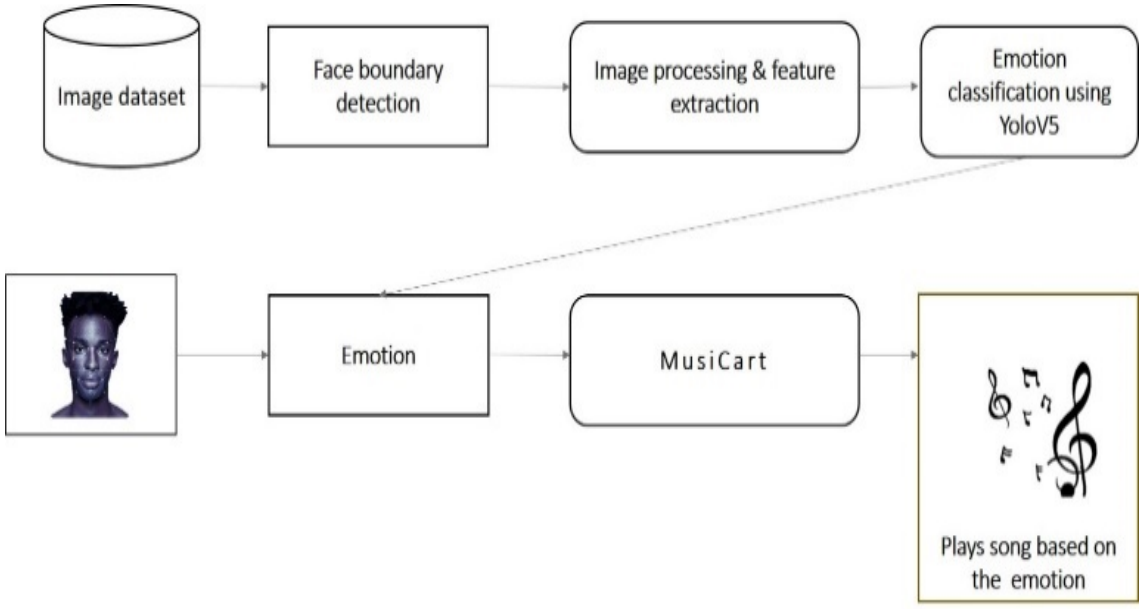


Fig 1 . System Architecture

IV. FACE EMOTION RECOGNITION

Image dataset

A dataset is a set of images. The dataset contains 3152 images for all emotions like angry, happy, neutral and sad in total. These datasets are taken from EPFL website. The EPFL-RLC dataset was captured using three static HD cameras in the EPFL Rolex Learning Center. The following list of images in **Fig 2** are downloaded dataset from EPFL website.

Fig 2. Image Dataset

Face boundary detection

Face detection is a computer technology that identifies human faces in digital images and is used in variety of applications. The psychological process by which humans locate and attend to faces in a visual scene is also referred to as face detection. This face bounding detection is done by bounding box. Bounding boxes are the most commonly used type of annotation in computer vision. Bounding boxes are rectangular boxes used to define the location of the target object. They can be determined by the x and y axis coordinates in the upper-left corner and the x and y axis coordinates in the lower-right corner of the rectangle. Bounding boxes are generally used in object detection and localization tasks. Face boundary is done using makesense.ai website. It is a free-to-use online tool for labelling photos. Image labelling is the process of identifying and marking various details in an image. It is useful when automating the process of generating meta data or making recommendations to users based on details in their images. Dataset is labeled for 3152 images. The **Fig 3** shows the bounding box of each image in the dataset created using makesense.ai site.

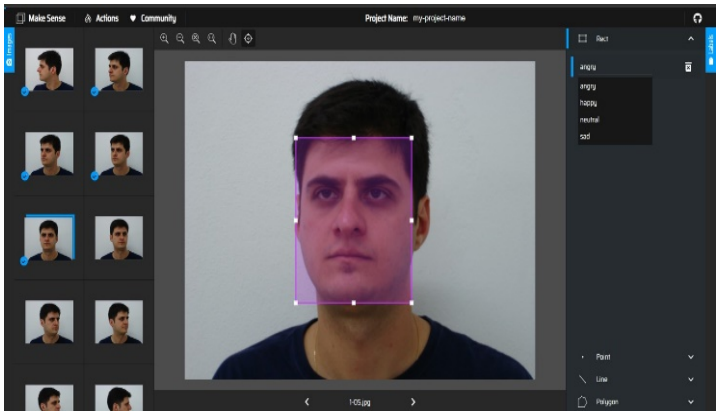


Fig 3. Face Bounding Box

Image processing and feature extraction

Image processing is a method to perform operations on an image, in order to get an enhanced image or to extract useful information from it. After face boundary detection, the labels are extracted from makesense.ai website. The label shown in **Fig 4** containing coordinates (x1, y1) and (x2, y2) or by one co-ordinate (x1, y1) and width (w) and height (h) of the bounding box.

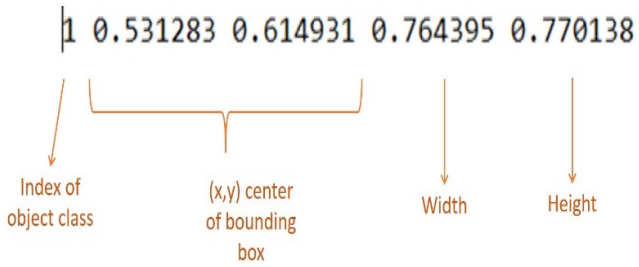


Fig 4. Bounding Box Showing Co-Ordinates X1, Y1, X2, Y2,Width and Height

V. CLASSIFICATION OF EMOTION USING YOLOV5

Four steps are involved in the classification of emotions using YOLOv5 are listed as follows.

Preparing dataset

The dataset contains images of various vehicles in varied face emotion. These images have been collected from the EPFL (Ecole Polytechnique Federale de Lausanne). EPFL-RLC (Rolex Learning Center) dataset was recorded in the EPFL Rolex Learning Center using three static HD cameras. Unlike most of the existing multi-camera datasets, the cameras fields of view are overlapping. Each camera has a resolution of 1920*1080 pixels and during the acquisition a frame rate of 60 frames per second was used. The images are from varied conditions and scenes. It contains 4 classes in total. They are Angry, Sad, Neutral, Happy. Labeled dataset are obtained from EPFL site and annotated them using tool makesensi.ai. Annotations and images are kept in the same directory. Then train, test, and validation txt files are generated. Best practice would be to keep 70% data in the training set, 20% in the validation set, and 10 % in the testing set.

Environment setup

The important thing is that will require PyTorch version ≥ 1.5, and Python version 3.7. The rest of the below dependencies in **Fig 5** are installed using pip/requirement.txt file.

```
# Base ----------------
matplotlib>=3.2.2
numpy>=1.18.5
opencv-python>=4.1.1
Pillow>=7.1.2
PyYAML>=5.3.1
requests>=2.23.0
scipy>=1.4.1  # Googl
torch>=1.7.0
torchvision>=0.8.1
tqdm>=4.41.0
protobuf<=3.20.1  # h

# Logging -------------
tensorboard>=2.4.1
# wandb
```

Fig 5. Requirements For Yolov5

Configure/modify files and directory structure should be as show in following **Fig. 6.**

To train the YOLOv5 model, need to perform some steps.

- Start with the cloning repository for YOLOv5.
- Modify YAML (Yet another Markup Language) file to describe the dataset parameters.

```
train: ../train/images/train/  # train images (relative to
'path')
val: ../train/images/val/  # val images (relative to 'path')

# Classes
nc: 4
# number of classes
names: ['angry','happy','neutral','sad']  # class names
```

Fig 6. Coco128 Yaml File

Training and inference

To train the YOLOv5, Glenn has proposed 4 versions [ ].

- YOLOv5-s which is a small version
- YOLOv5-m which is a medium version
- YOLOv5-l which is a large version
- YOLOv5-x which is an extra-large version

While training the model, YAML file is passed to YOLOv5-s model.

Move to the directory and use the following command to start training.

!python train.py –img 640 –batch 16 –epochs 100 –data coco128.yaml –weights yolov5s.pt

- **img:** size of the input image
- **batch:** batch size
- **epochs:** number of epochs
- **data:** YAML file which was created in previous step
- **cfg:** model selection YAML file.
- **weights:** weights file to apply transfer learning.
- **device:** to select the training device, “0” for GPU, and “cpu” for CPU.

This command will start the model training immediately. Trained the model for 700 epochs.

After the model has been trained. It will test its performance on validation images.

To run the model inference, following command is used.

! python detect.py –weights yolov5.pt –img 640 –conf 0.25 –source data/images

- **source:** input images directory or single image path or video path
- **weights:** trained model path
- **conf:** confidence threshold

This will process the input and store the output in inference directory.

Validation

As a result of trained model inference, **Fig 7** shows the emotion that was detected for the image.

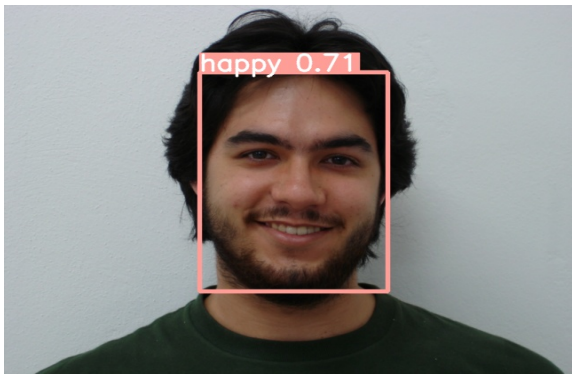


Fig 7. Detected Image

Once the training is completed, the model is saved in “weights” directory and the resulting matrix graph will be generated as following **Fig 8**. mAP (mean Average Precision) is an evaluation metric used in YOLO. It compares the ground-truth bounding

box with the detected box and returns score. The higher the score, the more precise the model's detections will be. The model is trained with mAP@0.5 is 0.95 as shown in **Fig 8** which indicates the model is trained well to detect the facial emotion.

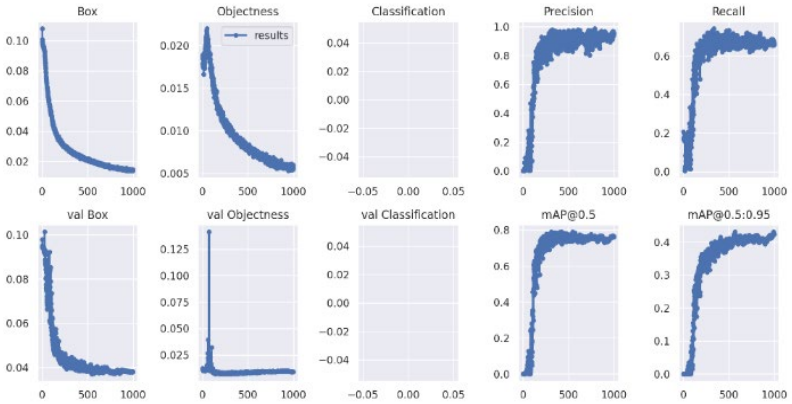


Fig 8. Resulting matrix graph

VI. MUSIC RECOMMENDATION SYSTEM

The login page contains a username and password for the users to login as shown in **Fig 9.**

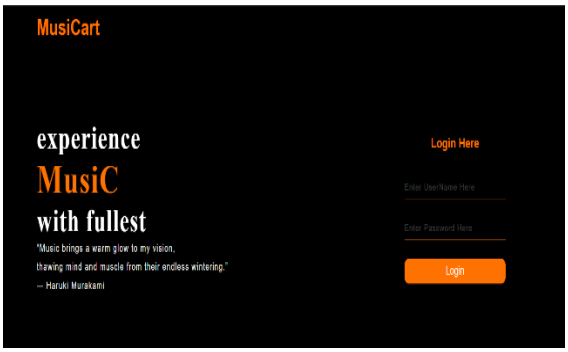


Fig 9. Login Page

The main page as shown in **Fig 10** contains all categories of songs. The user can listen to music they want. Songs are categorized according to the emotions (angry, sad, neutral, and happy).

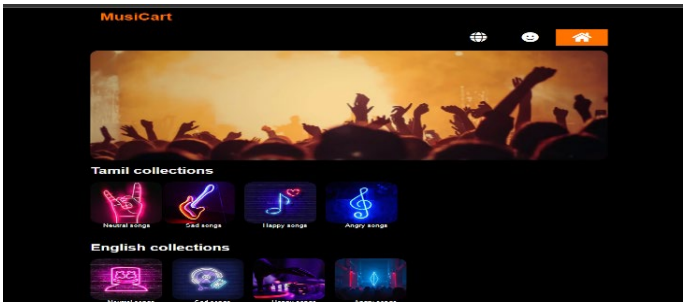


Fig 10. Main Page

The emotion page is build using the Flask API. It is an API of Python that allows us to build web applications. The user's face will be captured using a webcam. Emotion is detected using YOLOv5 algorithm and the detected image is displayed in MusiCart as shown in **Fig 11** Songs are displayed based on the emotion detected.

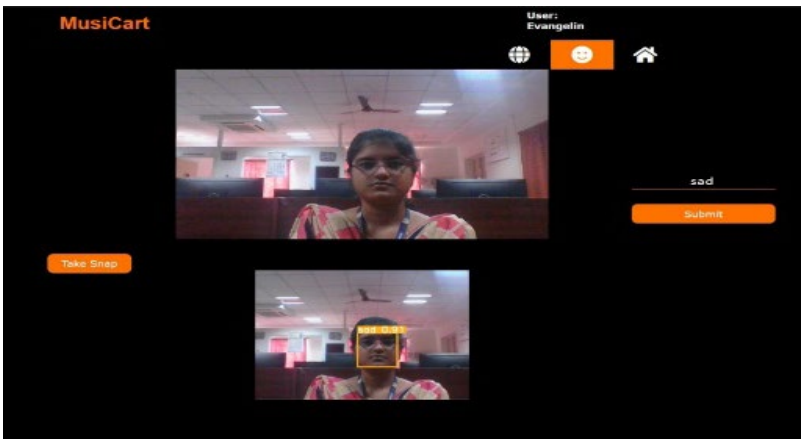
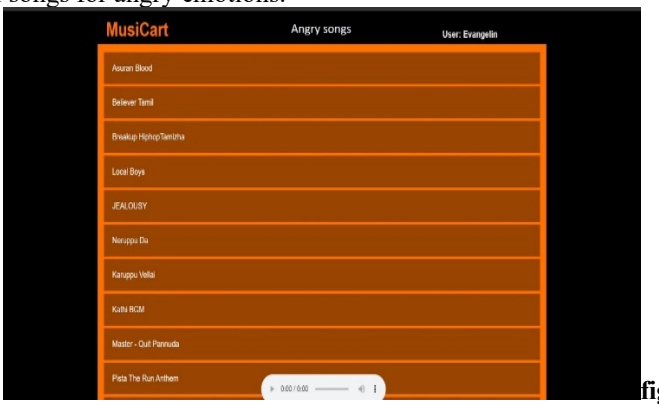


Fig 11. Detecting Face Emotion

The following **Fig 12** is a set of songs for angry emotions.



fig

Fig 12. Angry Songs

The following **Fig 13** is a set of songs to recommend for happy emotions based on user playlist category.

Fig 13 .Happy Songs

The following **Fig 14** is a set of songs for sad emotions based on user's playlist category.

Fig 14. Sad Songs

IV.CONCLUSION

The recommendation of music based on facial emotion using machine learning technique is intended to improve the interaction between the music system and the user because music helps to change the user's mood and it's a stress relief for certain people. The object recognition system can be used in surveillance systems, face identification, fault detection, character recognition and also used in face emotion. The goal of this project is to create a music recommendation system based on facial emotion. The point detection algorithm is used for feature extraction and YOLOv5 is used for training face emotion for promising results. Thus, system provided 0.95 accuracy on real face images. The emotion obtain are passed to music website created which shows a list of songs based on their emotion. The limitation of face emotion recognition detects person only when the user is directly facing the camera and it detects users when they are in bright surroundings. The problem is to improve this system to detect facial emotions of the people who are not directly facing the camera and it should also be able to detect them in the dull background by increasing the dataset. Music recommendation system can be further developed by implementing an algorithm for the music website to list top songs based on the user's listing list.

References

[1] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004,10934 (2020).

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, doi: 10.1109/cvpr.2016.91.

[3] W. Liu et al., "SSD: Single Shot MultiBox Detector," Lecture Notes in Computer Science, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[4] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, doi: 10.1109/cvpr.2017.690.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2014, doi: 10.1109/cvpr.2014.81.

[6] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, doi: 10.1109/cvpr42600.2020.01155.

[7] Park, Jongchan, Sanghyun Woo, Joon-Young Lee, and In So Kweon. "Bam: Bottleneck attention module." arXiv preprint arXiv:1807.06514 (2018).

[8] Ferwerda, Bruce, and Markus Schedl. "Enhancing Music Recommender Systems with Personality Information and Emotional States: A Proposal." In Umap workshops. 2014.

[9] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks," Mathematical Problems in Engineering, vol. 2019, pp. 1–9, Dec. 2019, doi: 10.1155/2019/2593036.

[10] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models," Sensors, vol. 21, no. 4, p. 1249, Feb. 2021, doi: 10.3390/s21041249.

[11] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, doi: 10.1109/icassp.2018.8462677.

[12] A. Chowanda, R. Sutoyo, Meiliana, and S. Tanachutiwat, "Exploring Text-based Emotions Recognition Machine Learning Techniques on Social Media Conversation," Procedia Computer Science, vol. 179, pp. 821–828, 2021, doi: 10.1016/j.procs.2021.01.099.

[13] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding," Frontiers in Computer Science, vol. 2, May 2020, doi: 10.3389/fcomp.2020.00014.

[14] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Dec. 2016, doi: 10.1109/apsipa.2016.7820699.

[15] D. Tang, P. Kuppens, L. Geurts, and T. van Waterschoot, "End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, May 2021, doi: 10.1186/s13636-021-00208-5.

[16] T.-W. Sun, "End-to-End Speech Emotion Recognition with Gender Information," IEEE Access, vol. 8, pp. 152423–152438, 2020, doi: 10.1109/access.2020.3017462.

[17] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," IEEE Access, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/access.2019.2936124.

[18] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576–1590, Jun. 2018, doi: 10.1109/tmm.2017.2766843.

[19] R. Ranjan et al., "A Fast and Accurate System for Face Detection, Identification, and Verification," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 1, no. 2, pp. 82–96, Apr. 2019, doi: 10.1109/tbiom.2019.2908436.

[20] W.-Y. Hsu and W.-Y. Lin, "Ratio-and-Scale-Aware YOLO for Pedestrian Detection," IEEE Transactions on Image Processing, vol. 30, pp. 934–947, 2021, doi: 10.1109/tip.2020.3039574.