Identification of a Deep Learning Technique for Facial Emotional Datasets for Song Recommendation System

Dr. Albert J Mayan [1], Dr. Mary Posonia [1], Hariharan B P [1], Jagadeesh N and Mohnish Devaraj

Department of Computer Science Engineering
(of Affiliation)
Sathyabama Institute of Science and Technology
(of Affiliation)
Chennai, India.
albertmayan@gmail.com [2]

Department of Computer Science Engineering
(of Affiliation)
Sathyabama Institute of Science and Technology
(of Affiliation)
Chennai, India.
maryposonia.cse@sathyabama.ac.in

Department of Computer Science Engineering
(of Affiliation)
Sathyabama Institute of Science and Technology
(of Affiliation)
Chennai, India.
bphariharan1301@gmail.com

Department of Computer Science Engineering
(of Affiliation)
Sathyabama Institute of Science and Technology
(of Affiliation)
Chennai, India.
jagadeeshshirs6453@gmail.com

Department of Computer Science Engineering
(of Affiliation)
Sathyabama Institute of Science and Technology
(of Affiliation)
Chennai, India.
mohnish.devaraj312002@gmail.com

Abstract. An Android application called "SONG Recommender System Using Emotion Detection" allows usersto play music in accordance with their mood (emotion). The user can choose their language and employ emotion recognition using this app. All the prior research employed various algorithms, such as fisher faces, SVM, K-nearest algorithms, etc., however those methods were not always effective. Recent research recommended combining emotion recognition (ER) with deep learning (DL) applications to lessen the amount of manual work required from the user to search for and listen to songs. Convolutional neural networks (CNN) can be used to determine users' emotional recognition using the FER2013 dataset, according to this systematic review.

1 Introduction

By identifying and capturing the user's emotions in real time, this notion suggests music to the user. We presented a method to categorize various types of music into distinct moods, such as joyful, sad, irritated, etc.
Existing techniques that were employing collaborative techniques that would use user's data from prior sessions to select music. However, these strategies take a lot of manual work.
Additionally, the branch of research known as Computer Vision (CV) enables computers to learn by employing various strategies and methodologies in order to capture what is present in an image or a video. Currently, computer vision has a diverse array of applications, such as recognizing faces, autonomous vehicles, medical analysis, and more. Today, we will delve into the exciting field of detecting emotions through facial expressions, a noteworthy use of CV. By seeing the facial expressions, the model can identify and describe your sentiment.

2 METHODS

This study is solely focused on determining the best deep learning (DL) algorithm for emotional recognition that provides almost perfect accuracy. The necessary data was collected using a systematic review process, and quantitative methods were used to analyse and filter the datacollected in ac-

cordance with the needs of the study.

In this work, the following critical methods have been applied: -

1. Facial Detection — The capability to detect a face in any given image or frame using facial detection. The results are the detected faces' bounding box coordinates.

2. Facial Recognition—To determine which faces belong to the same person, compare various faces together. Comparing facial embedding vectors allows for this.

3. Emotion Detection — identifying the various facial expressions as pleased, angry, sad, neutral, or surprised.

CREATING RESEARCH QUESTIONS (RQ):

The following table1 shows the developed RQs according to the author's motivation for this research.

S.No	RQs	Motivations
1	Which is the most suitable DL technique for emotion recognition?	Researchers have used various DL techniques but, we are yet to know which has the near perfect accuracy.
2	What is the most suitable freely available face dataset to train and test the developed DL model?	Researchers have used various datasets to train their particular model but, for our study we have chosen The FER2013 datasets from Kaggle which has a wide variety of data in it (50,000) and the test accuracy has Reached 63.2%.

3   LITERATURE REVIEW

Examining the techniques is carried out to understand their weaknesses and how to improve them. The text of a scholarly paper, known as a literature review, encompasses the latest research findings and key theoretical and meth-

odological advancements in a specific field. There is worldwide interest among students, researchers, engineers, and other professionals in the latent abilities of individuals to provide inputs to systems through various means. Facial expressions can give insight into a person's emotional state. In interpersonal communication, we often convey our emotions through hand gestures, facial expressions, and vocal tones. Many users have reported that creating and maintaining a large playlist takes a considerable amount of time. According to the report, the music player adapts to the user's mood and selects a song accordingly. To generate mood-based playlists, the application classifies audio files based on audio attributes.To evaluate the efficacy of the developed model, Support Vector Machine (SVM) was used to categorize features into five major universal emotions: anger, joy, surprise, sadness, and disgust.

Ajinkya Khobragade [2] suggested a song recommendation system utilizing Python and Tableau, but he desired for this system to operate in a real-time setting and to capture the alignment of the facial structure to be developed for the comfort of users.
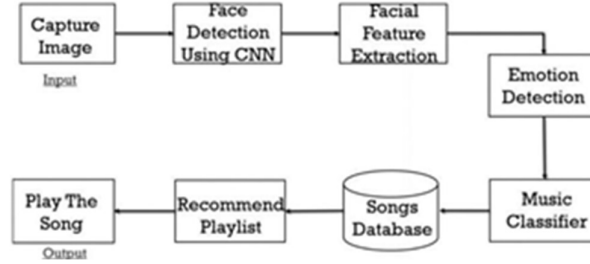

Fig 3.1: Yusuf's work flow

Yusuf Yaslan[12] and his colleagues proposed an emotion- based music recommendation system that utilizes wearable technology with physiological sensors for galvanic skin response (GSR) and photo plethysmo graphy (PPG) to detect the user's emotions. Emotions are a key element of human nature and play a vital role in our lives. This study uses multi- channel physiological signals to predict arousal and valence inorder to address the difficulty of emotion detection.
According to Ayush Guidel et al., facial expressions make it simple to determine a person's emotional state of mind at any one time.

4   PROPOSED SYSTEM OVERVIEW

The user-music player interaction in the suggested system is advantageous to us. This system's objective is to use the camera to acquire a facial image. The CNN then utilized to forecast the emotion using the captured visuals. A customized list of songs is created based on the detected response in the captured image. The objective of the system is to automatically generate a

music list that reflects the user's responsiveness.

Four modules make up the response-based music system:

• On-spot Capture: The system must accurately capture theperson response in this module.

• Face Recognition: Here, the person response will beused like data. The CNN is set up to analyze the pictorial characteristic of a person.

• Emotion Detection: In this stage, the user image's features are extracted in order to identify their emotions, and the algorithm then generates captions based on those emotions.

• Music Recommendation: By matching the user's emotions to the song's mood type, the recommendation module suggests songs to them

5   METHODOLOGY

The dataset was taken from Kaggle to create the CNN model. The dataset, FER2013, is split into combination of two different dataset, said to be training and testing dataset. There are total of 30,219 pics of data, in which 80%(24,176) is used for the training the model and 20%(6043) used for testing.

The collection contains 48x48 pixel grayscale portraits of data. In FER-2013, each data is assigned to a particular emotion among the set of five pre-loaded emotions such as neutral, happy, sad, angry, and surprised. Faces present in the image are automatically recognized in order to make sure it is approximately centred and occupies the same amount of area in each photograph. Google image searches for every emotion which were used to compile the dataset.

Fig 5.1 An Angry Face

When trained, the datasets with unequal distribution of emotions, facial ex-

pression recognition systems may excel in recognizing dominant emotions such as happy, sad, angry, neutral, and surprised, but fall short in accurately identifying less frequently represented emotions.

A more detailed view on how the algorithm works: -

```
        ┌──────────────┐
        │    START     │
        └──────────────┘
               │
               ▼
     ┌──────────────────┐
     │ Dataset Analysis │
     └──────────────────┘
               │
               ▼
     ┌──────────────┐        ┌──────────────────┐
     │    Image     │───────▶│  Creating CNN    │
     │Preprocessing │        │  architecture    │
     └──────────────┘        └──────────────────┘
               │                      │
               ▼                      │
     ┌──────────────┐◀────────────────┘
     │Model Training│
     └──────────────┘
               │
               ▼
   ┌────────────────────┐
   │Real Time Emotion   │
   │    Detection       │
   └────────────────────┘
               │
               ▼
   ┌────────────────────┐
   │       Song         │
   │  Recommendation    │
   └────────────────────┘
               │
               ▼
        ┌──────────────┐
        │    STOP       │
        └──────────────┘
```
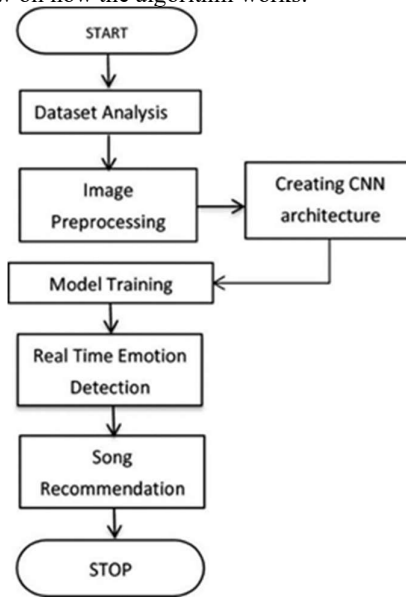
Fig: 5.2 Algorithm Flow Chart

This issue is typically addressed using the weighted-SoftMax loss strategy attempts to address this issue by assigning higher weights to the underrepresented classes. This way, the model is encouraged to pay more attention to the underrepresented classes and improve its accuracy in recognizing them. The weighted-SoftMax loss is calculated as the cross-entropy loss between the predicted probability distribution and the true distribution, where the weights are applied to the true distribution. We have applied the categorical cross-entropyloss function to handle missing and outlier values. A chosen loss function is used for each iteration to estimate the error value. Therefore, the utilized loss function called categorical cross-entropy to treat missing and outlier values.

Steps:

Step 1: The user gives the input, in the form of the captured image using the camera.

Step 2: The image is then analysed by the model and gets classified into a particular category of emotion.

Step 3: The data gets extracted and detected with the training datasets, which arefer-13 datasets.

Step 4: A particular customized list of songs(playlist) is chosen according to the user's emotion.

Step 5: The music is played depending on the emotion to of the user.

Face Detection:

One of the use of Computer Vision is facial detection. Process by which a specific algorithm is made and taught to identify faces. A video or image can be used to perform this detection in real time. This approach employs classifiers, algorithms designed to determine the presence [1] or absence of a face in an image. To boost accuracy, the classifiers undergo training with a substantial number of data to recognize faces. The technology OpenCV utilizes twodistinct types of classifiers, namely Local Binary Pattern (LBP) and Haar Cascades. In order to effectively detect faces, a Haar Cascade classifier is utilized. It is trained with pre-defined variable face data. Facial detection primarily focuses on locating the face within the frame by disregarding external disturbances and distractions. The technique is grounded in machine learning and utilizes a set of input files for training the cascade function. The Haar Wavelet approach is used to divide the image pixel into squares based on their functions [9]. The extraction of training data is achieved with a high level of accuracy through the utilization of machine learning techniques.

Emotion Detection:

The CNN architecture creates feature maps or activation maps using the Relu activation function by applying filters or feature detectors to the input image. These filters or feature detectors can identify edges, lines, curves, and other characteristics present in the picture. The feature maps are then subjected to pooling to maintain translation invariance, meaning that even small changes in the input should result in unchanged outputs. While pooling can be performed using minimum, maximum, or average, max-pooling is considered superior in performance compared to the other methods. Finally, the inputs are flattened before being fed into a deep neural network, which outputs the class of the object.
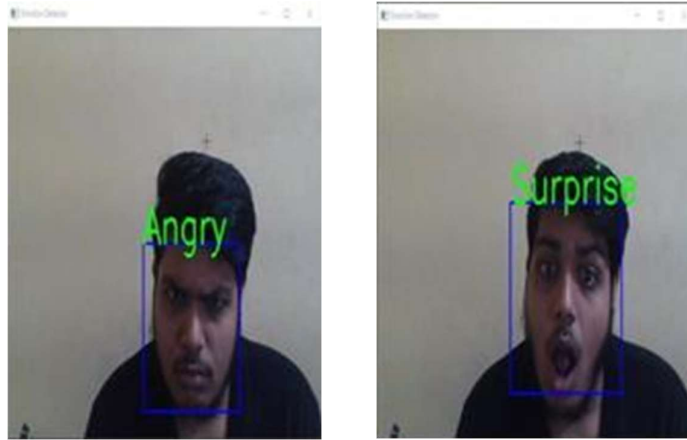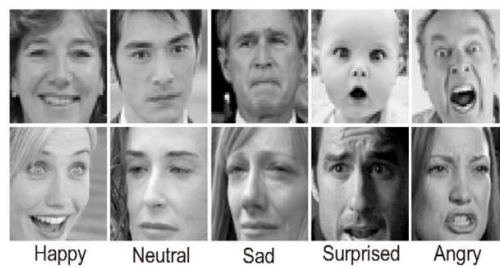
Fig: 5.3 Emotion Screen shots

Happy   Neutral   Sad   Surprised   Angry

Fig: 5.4 Dataset Image Samples

The Process of analysing the pictorial characteristic of a person.

1. Data Collection: Gather a dataset of labelled images of people. The dataset should be diverse and representative of the population you want to analyse, with a wide range of ages, genders, ethnicities, and body types.

2. Data Pre-processing: Prepare the dataset for training the CNN. This may involve resizing the images to a consistent size, normalizing pixel values, and splitting the dataset into training and validation sets.

3. Model Training: Train a CNN model on the pre-processed dataset. The CNN should consist of convolutional layers for feature extraction and pooling layers for down sampling, followed by fully connected layers for classification. The number and configuration of layers will depend on the specific CNN architecture you choose.

4. Feature Extraction: Once the CNN is trained, use it to extract features from new images of people. Pass the images through the trained CNN

and obtain the output from the desired layer(s) that captures relevant pictorial characteristics such as facial features, body shape, or clothing.

5. Feature Analysis: Analyse the extracted features using appropriate techniques. This may involve statistical analysis, visualization, or other methods to identify patterns, trends, or differences in the features across the dataset. You may also use domain-specific knowledge or external references to interpret the meaning of the features in the context of person characteristics.

6. Model Evaluation: Evaluate the performance of the CNN model in capturing pictorial characteristics of people. This may involve quantitative metrics such as accuracy, precision, recall, or F1 score, as well as qualitative assessments of the model's ability to accurately represent the pictorial characteristics of the person images.

7. Interpretation: Interpret the findings from the feature analysis and model evaluation to draw conclusions about the pictorial characteristics of the person images. Document and communicate the results of your analysis in a clear and concise manner, and consider the limitations and potential biases of the CNN model and dataset used.

The basic process involves feeding an input image into a CNN model, which then produces the results [10]. A CNN- trained model, with learned weights, is utilized for emotion detection. When a real-time photo is submitted by a user, it is processed by the trained CNN model, which assigns an emotion label to the image based on its prediction.
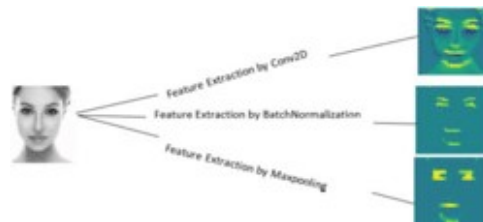

Fig: 5.5 Features Extracted for identification

6 RESULT & ANALYSIS

By default, this solution recognizes all faces in the image feed that exhibit emotions. The test accuracy for a straightforward 4-layer CNN was 63.2% after 50 iterations.
There were few overfitting during training due to the highcomplexity of transfer learning models and the comparatively modest size of the datasets. The ResNet transfer learning model soon over fit to the training data, with train

accuracy starting to flatten after only 25 epochs, even though 30% of dropout added in the last three layers.

The model comprises of a softmax output layer, a 1024-by- 1024-pixel FC layer, and three stages of convolutional and max-pooling layers. Filters of sizes 5x5, 4x4, and 5x5 are used in 32, 32, and 64 convolutional layers, respectively.

The kernels used by the max-pooling layers are 3x3 in size. As per the activation function, ReLU was used.

Additionally, batch norm was added at every layer and 30% dropout after the last FC layer to boost performance. Optimize the cross-entropy loss using a momentum of 0.3 to fine-tune the model. Initial learning rates of 0.2, 32, and 0.0001 are pre-determined, as are batch size and weight decay. If the validation accuracy does not increase, the learning rate is cut in half.

Epoch occurs only when the neural network processes an entire dataset at ONCE, which is both forward and backward.

7   CONCLUSIONS

The programme seeks to give operating system users a more user-friendly, versatile, and dependable emotion-based music system. People looking for music based on emotion and emotional behaviour would benefit from the Emotion-based Music Program. It could aid in reducing the amount of time spent searching for music, decreasing the amount of time wasted on computing, and ultimately improving the system's accuracy and efficiency. By employing a camera to capture the image, identifying the end-emotion, user's and providing a personalised playlist, it makes the end-job user's easier.

With the help of this initiative, machine learning technology will advance significantly. The purpose of an emotion-based music player is to organise the music according to the user's emotions, such as happiness or sadness.

8   FUTURE SCOPE

Even though this system is fully functional, there is still room for future development. The application can be altered in several ways to improve user experience overall and get better results. Some of these use an alternate approach based on extra emotions such as disgust and fear that are not included in the model.

The system's future plans include designing a mechanism that could aid in the treatment of individuals who are experiencing mental stress, anxiety, acute depression, and trauma through music therapy. Due to the current system's poor camera resolution and performance in extremely low light levels, there

is a chance to add some more functionality as a potential fix in the future.

In the future, this might be improved and added to iPods and other similar gadgets, making it much more practical.

9   LIMITATIONS

Due to the limited number of photos in the used image dataset, the system is currently unable to accurately record allemotions. For the classifier to produce accurate results, the image that it receives must be captured in a well-lit environment. The classifier can accurately estimate the user's sentiment if the image quality is at least higher than 320p.

10   ADVANTAGES OF PROPOSED SYSTEM

- Users do not need to select songs manually.
- No need of existing playlist.
- Users do not need to classify

References

[1] Preema J.S, Rajashree, Sahana M, Savitri H, Review on facial expression-based music player, International Journal of Engineering Research & Technology (IJERT), ISSN- 2278-0181, Volume 6, Issue 15, 2018.

[2] AYUSH Guidel, Birat Sapkota, Krishna Sapkota, Music recommendation by facial analysis, February 17, 2020.

[3] CH. sadhvika, Gutta.Abigna, P. Srinivas reddy, Emotion-based music recommendation system, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad; International Journal of Emerging Technologies and Innovative Research (JETIR) Volume 7, Issue 4, 2020.

[4] Julian Menezes .R, Albert Mayan .J, M. Breezely George," Development of a Functionality Testing Tool for Windows Phones", Indian Journal of Science and Technology,Vol:8,Issue:22,pp: 1-7, 2015.

[5] Vincent Tabora, Face detection using OpenCV withHaar Cascade Classifiers, Becominghuman.ai,2019.

[6] Zhuwei Qin, Fuxun Yu, Chenchen Liu, Xiang Chen. How convolutional neural networks see the world - A survey of convolutional neural network visualization methods. Mathematical Foundations of Computing, 2018.

[7] Frans Norden and Filip von Reis Marlevi, A Comparative Analysis of Machine Learning Algorithms in Binary Facial Expression Recognition, TRITA-EECS-EX- 2019:143.

[8] Ahmed Hamdy AlDeeb, Emotion- Based Music Player Emotion Detection from Live Camera, ResearchGate, 2019.

[9] Mary Posonia A, S. Vigneshwari, Albert Mayan J, D. Jamunarani,"Service Direct: Platform that Incorporates Service Providers and Consumers Directly",International Journal of Engineering and Advanced Technology (IJEAT) , Vol.8 ,No.6,pp.3301-3304,2019.

[10] John, A.M.,Prince Mary, S., Mohana Prasad, K., Naveena, M. and Laveti, N. (2022),'Authentication for cloud computing system through smartcard', Int. J. Cloud Computing, Vol. 11, Nos. 5/6, pp.518–528.

[11] Albert Mayan J, Karthikeyan S, Nikhil Chandak, Bharat Mundhra and Padmavathy J, "Facial attendance system technology using Microsoft Cognitive Services", International Journal of Engineering Systems Modelling and Simulation, Vol. 12, Nos. 2/3, pp. 180-187 ,2021

[12] Yusuf Yaslan, Deger Ayata, Mustafa E. Kamasak, Emotion Based Music Recommendation System Using Wearable Physiological Sensors, IEEE journal, Volume 64, Issue 2, May 2018.