Music Recommendation Using Facial Emotion Recognition

Rutik Patel
Dept. of Computer Science
MIT School of Engineering
Pune, Maharashtra
rutikpatel1872@gmail.com

Vivek Pusti
Dept. of Computer Science
MIT School of Engineering
Pune, Maharashtra
vivekpusti1999@gmail.com

Rishabh Singh
Dept. of Computer Science
MIT School of Engineering
Pune, Maharashtra
rishabh15112001@gmail.com

Umang Patel
Dept. of Computer Science
MIT School of Engineering
Pune, Maharashtra
umangpatel.830@gmail.com

Nagesh Jadhav
Dept. of Computer Science
MIT School of Engineering
Pune, Maharashtra
nagesh.jadhav@mituniversity.edu.in

***Abstract*—Face recognition technologies have attracted a lot of attention because of their ever-increasing demand in various sectors. It is used in a variety of fields such as security systems, digital video processing, recommendation systems, and other technological advancements. Additionally, music is a form of art, known for having a close connection with one's emotions. By comparison, this paper focuses on building an effective music recommendation system that determines user emotions using the Face Recognition technique. The classification of emotion based on facial features can be done using CNN and Computer Vision. Convolutional Neural Network (CNN) is one of the most popular tools used for facial expression recognition. Based on the classified emotion, music recommendation is performed from music datasets.**

***Keywords—Machine learning, Deep learning. Convolutional neural network OpenCV, Face Recognition*.**

I. Introduction

In recent years, recommendation algorithms have become a desirable topic. Machine learning approaches have been increasingly popular in recent years for recommendation systems. The field of music is one of the fields in which recommendation algorithms can be used. Music has long been linked to human feelings. A recommendation system based on emotion recognition could be a useful tool for the application at hand. Although detecting human emotion is difficult, advances in FER (Facial emotion recognition) have aided in the process. FER may be done properly with the help of technologies like CNN[1].

In this research article, we will describe how we will train our model to predict music from facial expressions using the FER2013 dataset, which is a popular dataset used in Kaggle competitions and a music dataset. We'll also use a convolutional neural network (CNN) to create a model that extracts features from the dataset's input photos. CNN and OpenCV are used to align faces on input images, while Keras is utilized to implement them.

.

II. Methodology

The description of the strategies used to implement a learning model will be presented in this section. CNN is used to build a computational model that assists in the classification of facial features, allowing for the classification of emotions into categories such as happy, sad, angry, and neutral.

A. Dataset Description

The dataset used to train the model came from the FER2013 Kaggle Facial Recognition Challenge[2]. The data consists of grayscale images of faces at a resolution of 48x48 pixels. The faces have been automatically registered such that they are more or less centred in each image and take up roughly the same amount of area. The goal is to categorize each face into one of seven categories based on the emotion expressed in the facial expression

(0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

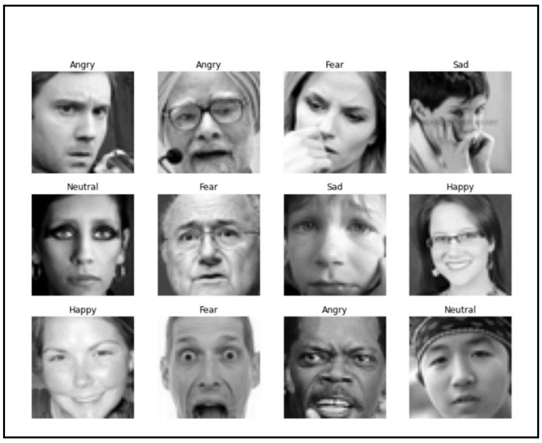

Figure 2.1: Sample images from FER2013.

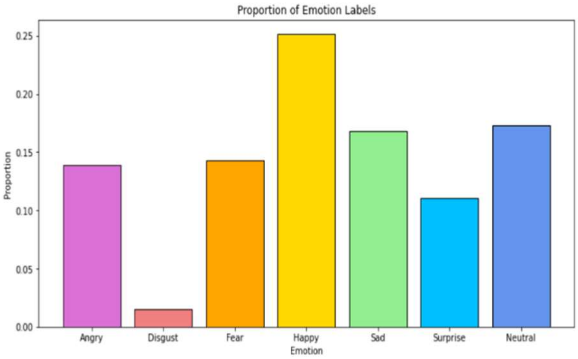

Fig 2.2: Distribution of sample images across different classes of emotion.

B. Data Pre-processing

Cleaning data takes place at this stage, which involves removing noise from the data. The phases involved in preparing data in the form of images are listed below.

***Face detection**:* Face detection is a technique for determining whether or not an image contains a human face. It is a subset of object class detection that examines an image for the presence of a face.

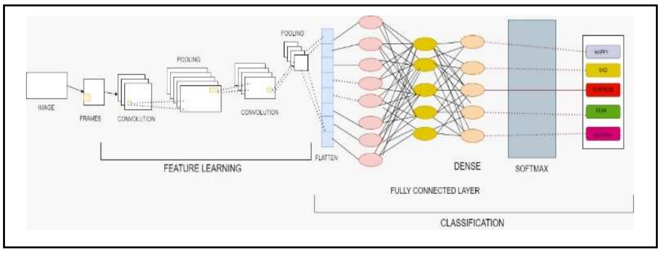

Fig 2.3: Overview of Convolutional Neural Network

Open CV is used for the purpose of face detection, which would further be used for predicting the emotion using the trained model.

***Balancing the dataset**:* Imbalanced data is a term used to describe datasets in which the target class has an unequal distribution of observations. This phase is critical since any dataset must have almost equal numbers of examples of each class. Oversampling or under sampling are two strategies that can be used to achieve balancing. As observed from Fig 2.2, FER2013 dataset is imbalanced. Oversampling techniques can be applied to make the data balanced[3].

***Normalization**:* The data is normalized before feeding it to the CNN model. As the given data is in form of pixel values that lie in the range (0, 255), the normalization can be done simply by dividing the values by 255. The resulting values lie in the range of (0, 1) which is helpful for convergence of gradient descents.

C. Model Description

CNN's have usually been a popular choice for working with image datasets. A CNN uses a system much like a multilayer perceptron that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer, and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers, and normalization layers. The removal of limitations and increase in efficiency for image processing results in a system that is far more effective, and simpler to train limited for image processing and natural language processing[4].

i. Convolutional Layers: These layers are used to apply filters in order to extract features from the image. The number of filters can be passed as a parameter for training the model. Filters are composed of kernels whose size is also passed on as a parameter. Each feature detector slides across the entire image generating a set of the feature maps. This results in a decrease in dimensionality. Padding is the technique used to preserve the size of feature maps.

ii. Pooling Layers: After the convolutional layers, the information is passed on to a pooling layer which is responsible for reducing the dimensionality. This helps in reducing the number of parameters, thus saving a lot of computational resources. The widely used approaches in pooling comprise Max Pooling and Average Pooling however, in the proposed model we have used Max Pooling.

iii. Dense Layers: After a series of convolutional and pooling layers, the data is passed on to dense layers. Dense layers require the data to be flattened. The choice of activation function for this layer is SoftMax.

D. Music Recommendation Module

The Music can be recommended in the form of songs that are already classified based on the emotion labels. However, the interpretation of a particular mood or emotion of a song might vary from person to person.

A pre-classified dataset containing songs that are labelled according to an associated emotion can be used for recommending songs from the current emotion of the user. The user's emotion could be obtained by taking a photo sample of the user and then using the CNN model to classify the emotion.[5]

CONCLUSION AND FUTURE SCOPE

The proposed music separation module works very well in recognizing facial features and emotions based on which music recommendation takes place, and thus, the model reduces the user's efforts to classify songs manually by successfully mapping the user's feelings to the correct song class.

The system is still unable to record all emotions properly due to the limited availability of images in the image database used. Labelling the songs to a particular emotion type could also be improved. We have decided to address these issues in future work.

REFERENCES

[1] Naskar, Avijit & Mandal, Akrito & Ghosh, Anirudha & Saha, M & Sufian, A.. (2019). Convolutional Neural Networks. 10.13140/RG.2.2.23743.05282.

[2] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, "A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets, and Future Directions," in IEEE Access, vol. 9, pp. 165806-165840, 2021, DOI: 10.1109/ACCESS.2021.3131733.

[3] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," 2017 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), 2017, pp. 79-85, doi: 10.1109/ICACCI.2017.8125820.

[4] Srushti S Yadahalli, Shambhavi Rege, Sukanya Kulkarni, "Facial Micro Expression Detection Using Deep Learning Architecture", *Smart Electronics and Communication (ICOSEC) 2020 International Conference on*, pp. 167-171, 2020.

[5] Shlok Gilda, Husain Zafar, Chintan Soni, Kshitija Waghurdekar, "Smart music player integrating facial emotion recognition and music mood recommendation", *Wireless Communications Signal Processing and Networking (WiSPNET) 2017 International Conference on*, pp. 154-158, 2017