ACTIVITY RECOMMENDATION SYSTEM BASED ON FACIAL AND SPEECH EMOTION RECOGNITION USING CNN

Rohit Bari[1], Vaishnavi Belsare[2], Tanmay Ghare[3], Ritika Sharma[4], Prof. Umesh Nanavare[5]

[1,2,3,4]Students of Smt. Kashibai Navale College of Engineering, Pune
[5]Professor of Smt. Kashibai Navale College of Engineering, Pune

---

Abstract

Emotion recognition is the process of identifying human emotion. People vary widely in their accuracy at recognizing the emotions of others. Use of technology to help people with emotion recognition is a relatively nascent research area. A Convolutional Neural Network (CNN) is a Deep Learning algorithm which can take in an input image/voice, assign importance (learnable weights and biases) to various aspects/objects in the image/voice and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics. Facial Emotion Recognition (FER) is the technology that analyses facial expressions from both static images and videos in order to reveal information on one's emotional state. Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch.

Key Words: Facial emotion recognition (FER), Speech emotion recognition (SER), Convolutional Neural Networks (CNN)

1. INTRODUCTION

Facial emotion recognition is the process of detecting human emotions from facial expressions. The human brain recognizes emotions automatically, and software has now been developed that can recognize emotions as well. This technology is becoming more accurate all the time, and will eventually be able to read emotions as well as our brains do. AI can detect emotions by learning what each facial expression means and applying that knowledge to the new information presented to it.

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.

Convolutional neural network (CNN) is the most popular way of analysing images. CNN is different from a multi-layer perceptron (MLP) as they have hidden layers, called convolutional layers. the conventional CNN network module is used to extract primary expressional vector (EV). The expressional vector (EV) is generated by tracking down relevant facial points of importance. EV is directly related to changes in expression. The EV is obtained using a basic perceptron unit applied on a background-removed face image. Each of the convolutional layers receives the input data (or image), transforms it, and then outputs it to the next level.

The goal of the project is to build an Activity Recommendation System based on Convolution Neural Networks to suggest mood enhancing activities to counter anxiety and depression by examining their facial expressions and voice intensities.

2. LITERATURE SURVEY

[1] Face Expression Recognition Based on Optimized Convolutional Neural Network - This paper optimizes the structure of the facial expression recognition network model based on VGGNet16 and improves the model performance. Adopting the optimization strategy of Batch Normalization, Cross-Entropy Loss Function, Stochastic Deactivation, and Data Enhancement Combination, the network model improves the generalization capability and speeds up the convergence of network training, then increases the ability to identify and classify facial expression recognition.

[2] A Comprehensive Review of Speech Emotion Recognition Systems - Two classification algorithms are used to recognize emotions, traditional classifiers, and deep learning classifiers, after the extraction of features. Even if there is much work done using traditional

techniques, the turning point in SER is deep learning techniques. Although SER has come far ahead than it was a decade ago, there are still several challenges to work on. Some of them are highlighted in this paper. The system needs more robust algorithms to improve the performance so that the accuracy rates increase and thrive on finding an appropriate set of features and efficient classification techniques to enhance the HCI to a greater extend.

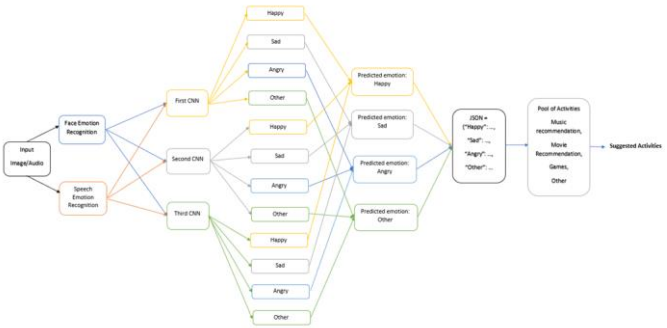
[3] Emotional Detection and Music Recommendation System based on User Facial Expression - The system has been able to grab the images of the user and appropriately update its classifier and training dataset. The system was designed using the facial landmarks scheme and was tested under various scenarios for the result that would be obtained. It is seen that the classifier has an accuracy of more than 80 percent for most of the test cases, which is pretty good accuracy in terms of emotion classification. It can also be seen that the classifier can accurately predict the expression of the user in a real-time scenario when tested live for a user.

[4] A computer vision-based image processing system for depression detection among students for counseling - This study was undertaken for finding out the level of depression in five different videos of college students. The presence of 'Happy', 'Neutral, - (positive emotion) and 'Contempt' and 'Digust'- (Negative emotion) facial features, which are found prominent in depression videos were found out and analysed. The dataset for training and testing was captured separately and the facial features of the same were classified using a Support Vector Machine classifier. The amount of the positive and negative emotions in each video was analysed and the videos were predicted as videos with 'High Depression', 'Mild Depression' or 'Low Depression'. The classifier predicted the outcomes with a maximum accuracy of 64.38% accuracy.

[5] Facial Expression Detection by Combining Deep Learning Neural Networks - In this paper, we have described a system for FER, which processes static images or video captures. Facial landmarks were extracted and analyzed for each person in each image to determine emotion. The system used a weighted average of outputs from three CNNS to predict emotions, the results being improved compared to those of each network taken individually. Unfortunately, the system could not be used with real time videos (at least 24 frames per second), as preprocessing to obtain the landmarks images is slow, taking up to 400ms per face. The system performs well over our dataset, which has images of males and females, of various ages and ethnicities. Reduction of redundant information through the conversion of source images into black and white representations of facial landmarks managed to reduce resource and time usage, but unfortunately came with an increase in incorrect results. In the future, we want to reduce processing times and achieve real time FER on videos. Also, we need to improve the accuracy of distinguishing the emotion of sadness.

[6] Emousic: Emotion & Activity based Music Player using Machine Learning - This paper presents Emousic, a new way of personalizing songs playlist by using machine learning techniques. Our solution works and gives better user preferable playlist. Due to the large number of classes, the performance of Random Forest classifier is better than Decision tree algorithms. Since the experiment was performed on a small dataset and limited number of features, it still can be improved by adding more features like age, weather etc. More number of attributes will improvise decision making and prediction of song. Each user has it's own preferences about what kind of song is to be played for corresponding mood. for e.g., some users listen sad songs when they are sad while some may prefer happy songs to change their mood. Collecting this data from every user can help us build better user specific radio application. Implementing this prototype in current music applications can provide better music experience to user.

3. PROPOSED ALGORITHM



a. Datasets

The system uses two large Kaggle datasets for processing.

i. Facial expression recognition dataset that includes 71,774 various male and female face images classified into 7 different groups of emotions namely happy, sad, angry, disgusting, fear, neutral and surprise.
ii. Speech emotion recognition includes audio-only files (16bit, 48kHz .wav) from the RAVDESS. This dataset contains 1440 files: 60 trials per actor x 24 actors = 1440. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

b. Input

It takes input in image or audio format using internal or external devices including webcam or microphone.

c. FER and SER

The obtained inputs are passed to the Face Expression Recognition (FER) or Speech Emotion Recognition (SER) algorithms respectively.

d. CNN

Then, it is passed to the three layers of Convolution Neural Networks architecture for enhancing efficiency and accuracy of the algorithms for predicting the user's emotion.

Three main types of layers are used to build the architecture:

i. Convolutional Layer - To transform the input image and extract features.
ii. Pooling Layer - To reduce the size of the input image so that it speeds up the computation of network.
iii. Fully-Connected Layer - To flatten the outputs from pooling layer and connect to every activation unit.

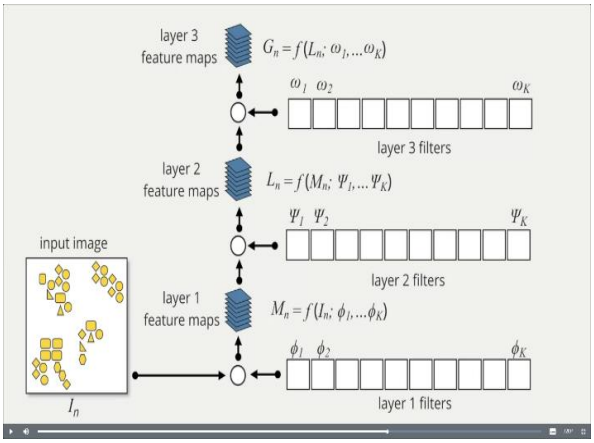
e. Activity Mapping

The predicted emotion from the trained model is mapped to the pool of activities stored in JSON.

f. Output (Activity Suggestion)

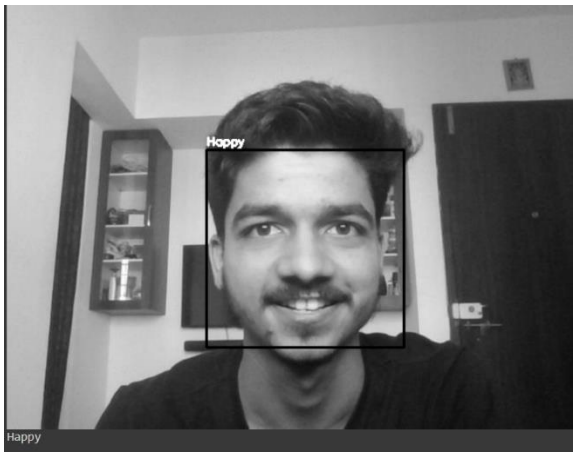
Finally, the appropriate activities are picked and suggested to the end user.

g. Mathematical Model

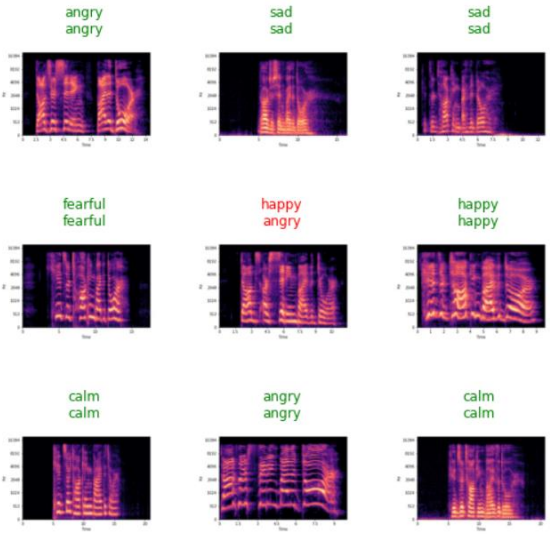


4. SIMULATION RESULTS

a. Predicted emotion by face expression recognition:



b. Predicted emotions using speech recognition:



5. CONCLUSION AND FUTURE WORK

We have described a system for FER, which processes static images or video captures. Facial landmarks were extracted and analyzed for each person in each image to determine emotion. The system used a weighted average of outputs from three CNNs to predict emotions, the results being improved compared to those of each network taken individually.

The system performs well over our dataset, which has images of males and females, of various ages and ethnicities. Reduction of redundant information through the conversion of source images into black and white representations of facial landmarks managed to reduce resource and time usage.

The system has successfully been able to capture the emotion of a user. It has been tested in a real-time environment for this predicate. However, it has to be tested in different lighting conditions to determine the robustness of the developed system.

ACKNOWLEDGEMENT

It is a matter of great pleasure for us to submit this report on "Activity Recommendation System Based on Facial and Speech Emotion Recognition using CNN" as a part of curriculum for award of "Bachelor's in Engineering (Computer)" degree of Savitribai Phule Pune University. Firstly, we would like to express my gratitude to my guide **Prof. Umesh Nanavare,** for his inspiration, adroit guidance, constant supervision, direction and discussion in successful completion of dissertation. We are also thankful to him for his support in providing their port format, making flexible time schedules for assignments, test, and lectures to all students during entire year. We are grateful to Head of Department **Dr**. **R. H. Borhade**, for his valuable support and guidance.

We are thankful to my **Principal Dr. A. V. Deshpande** and to all our staff members who encouraged us to do this, we also extend our thanks to all our colleagues those who have helped us directly or indirectly in completion of this project.

REFERENCES

[1] Wang – "Face expression Recognition Based on Optimized Convolutional Neural Network"

[2] Taiba Majid Wani ,Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi – " A comprehensive Review of speech emotion recognition system"

[3] Namboodiri Sandhya Parameswaram , D Venkatraman - "A computer vision based image processing system for depression detection among students for counselling"

[4] Jagannath Aghav – "Emousic : Emotion and activity - based music player Using Machine Learning"

[5] Alexandru Costache, Dan Popescu - "Facial Expression Detection by Combining Deep Learning Neural Networks"

[6] S Metilda Flourence, M Uma - "Emotional Detection and Music Recommendation System Based on User Facial Expression"

[7] F. Kong - "Facial Expression Recognition Method based on Deep Convolutional Neural Network Combined with Improved LBP Features"

[8] J. Hook, F. Noroozi, O. Toygar, and G. Anbarjafari - ''Automatic speech based emotion recognition using paralinguistics features''