

# Music Recommendation System Based On Facial Expression And Speech

Mrunmayee Shewale<sup>1</sup>, Sahil Sinha<sup>2</sup>, Prof. Satyajit Sirsat<sup>3</sup>

Department of Computer Engineering,  
Nutan Maharashtra Institute of Engineering and  
Technology, Talegaon Dabhade, Pune

**Abstract** - In recent years, with the development and use of big data, deep learning has begun to attract more and more attention. Convolutional neural network, a deep learning neural network, plays an important role in facial image recognition. This paper combines convolutional neural networks' knowledge of micro interpretation technology with an automatic music recognition algorithm to create patterns that recognize micro faces, speak, and recommend music based on your mood. The facial micro expression recognition model developed in this article uses FER 2013 and the recognition rate is 62.1%. After determining the similarity, the content-based music recognition algorithm was used to extract the feature vector of the song, and the cosine similarity algorithm was used for music recognition. This research helps improve the effectiveness of music recognition, and related results can also be applied to the use of music recognition in areas such as emotion regulation. Keywords: deep learning, face macro recognition, CNN, FER2013, CB, music recommendation algorithm

## I. Introduction

With the information age, deep learning is widely used in image recognition, image processing and especially face recognition. Facial recognition has become a research hotspot in the field of human computer interaction, but there are still limitations in the use of image processing. Image research mostly focuses on increasing identification accuracy, but the information in the image has no benefit in the sec and process, that is, in the production process and in real life, and the image data has not been processed and used effectively [1]. This article uses deep learning to design and train a neural network based cognitive model. Image processing results are combined with the music recommendation algorithm to recommend mind enhancing music according to the person's decision. Music files are created by accessing and writing to playlists on major music websites. Various applications of image processing have been expanded accordingly.

## II. Micro-expression Recognition

Basic Steps for Micro-expression Recognition

The basic process of facial micro-expression recognition is as follows:

- Obtain micro facial expression images of human faces and preprocess the images;
- Perform micro-expression detection and related feature extraction;
- Perform micro-expression classification.

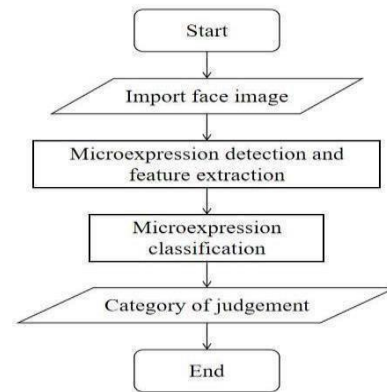


Figure 1: Flow Chart of Micro-Expression Recognition

## Micro-expressions and Feature Extraction

[1-3] An important step in the facial micro interpretive recognition system is the prioritization of facial images. Due to the influence of various factors, the quality of the input image (size, pixel, grayscale value, etc.) is not uniform, and the image cannot be directly used to bite the front of the face, which affects cognitive performance. Therefore, image preprocessing has a structure and integration that can eliminate the influence of size, body, light and shadow, and other aspects of cognitive processing. [4] Reduce the impact of irrelevant information and demonstrate the value of micro expressions. Histogram equalization, median filtering, grayscale stretching, homomorphic filtering, nearest neighbor algorithm, bilge linear interpolation, etc. [5] There are many methods available before. Different micro expression recognition systems require different image and recognition algorithms as well as different preprocessing methods. [6] This paper uses FER2013 data from the facial micro expression set to lift the image before processing expression recognition algorithms can make up for these shortcomings. [7] What this article uses is a feature extraction algorithm based on a convolutional neural network. Convolutional neural networks can represent learning, and can perform translation-invariant classification of input information according to the hierarchical structure of the input information.

## Convolutional Neural Network

CNN neural network refers to the convolutional neural network in which many image images can be extracted and learned. It is similar to other neural networks in that it uses forward propagation to input data and unpack the hidden layer; The backpropagation loss rate changes the parameters of the hidden layer. Activation of the latent process is based on the activation function indicating nonlinearity, which allows the neural network to arbitrarily predict a nonlinear function. [8] The difference between CNN and other neural networks is that it is a deep neural network. After several convolutional layers and sink layers are connected alternately, a fully connected layer is connected, thus changing the hidden layer in the network. However, unless the number of layers and learning increases, the

risk of compromise during training will increase and the ability of the pattern will deteriorate. [9] Important parts exist in CNN: one is the convolution pooling layer, and the extraction of image features is completed in this part; output is classified by the fully connected layer is the other part.

Convolutional layer: several convolutional units make it, and the parameters of each convolutional unit are optimized by back propagation algorithm. The convolution layer performs a convolution operation on the original image and a specific feature filter. During multiple convolution processes, each operation uses a different filter to map different features. Pooling layer: Pooling is a form of down sampling. Through the function of the pooling layer, the space size of the data will be reduced continuously, and the number of Measurements and calculations will be reduced. Overcompliance can also be managed to some extent.[10] Full connection method: After flattening the previous results, each node is connected to all nodes in the previous layer used to collect previous features.

**Model Design**

According to the value of the training data and the classification of the face micro interpretation function, a small 8layer CNN model was built, including 4 convolutional layers, 2 layers and 2 full connections. layers. The design is based on the development of AlexNet.[11] It uses the convolution process of two convolutional layers to perform reduction followed by pooling process to improve the extraction ability of the CNN network and reduce the learning speed of the model. After this layer, [12] Relu is added as the activation function and finally the results are put into the classification result obtained from two fully connected layers and the Softmax classifier.

**Model Structure :**

The model of the facial micro expression CNN model developed in this article is shown in Figure 2. Among them, 48\*48 means the network input is 48\* gray scale image data 48 pixels, 64 and 48 . 48 means the first convolution layer It means there are 6 - 4 feature map data with size 42\*42 pixels. The results of mixing and pooling are also very informative.

The last two layers are fully connected that the number of neurons in this layer is 5-12 and 7, respectively. The model parameter settings are shown in

Table

Table 1. CNN Model Parameter Settings

Layer Type	Core Size	Output
Input Layer		48*48

Convolutional Layer 1	64&3*3	64&42*42
Convolutional Layer 2	128&3*3	128&36*36
Pooling Layer 1	128&2*2	128&18*18
Convolutional Layer 3	256&3*3	256&12*12
Convolutional Layer 4	256&3*3	256&6*6
Pooling Layer 2	256&2*2	256&3*3
Fully Connected Layer 1		512
Fully Connected Layer 2		7

**Data Set**

This model uses FER2013 data for model training. FER2013 dataset contains 35886 face maps. These include 28,708 maps, 3,589 public test images and 3,589 private test image s. Each image consists of gray images with a size of 48\*48. The non text avatar has a total of 7 characters corresponding to the numbers 06. Specific expressions for Chinese and English texts are as follows: 0 Anger; 0 Anger; 0 Anger; 0 Anger; 1 Fear; 2 Fear. ; 3 Happy; 4 sad; 5 surprises; 6 neutral [8] . Among these, the happy heart is the most.

**Model Training**

Since CNN is a feedforward neural network, the training of CNN is divided into two tasks: forward propagation and backpropagation [9]. During forward propagation, each neuron in each convolution kernel of the convolution layer connects to local and local viewports of the input feature map of the forward layer and performs a convolution operation to extract features from it. After the error is added, the result is used by the output of the activation function to generate the neurons of the current layer. These neurons execute the current process.

Speech recognition is the process of converting spoken words into text. The goal is to train the CNN model to recognize and distinguish voices so that the system can recognize and understand human speech. Application of CNN to audio data provides good results as it can learn and extract key features of spectrograms and other audio representations. Pre-recorded files:

Pre-recorded recorded files containing sound samples of various sounds. Convert the audio file to the appropriate format and preview the file to determine the audio length and sample size. So we need to convert the audio into a visual representation that CNN can do. Commonly used representations include Mel Frequency

Cepstrum Coefficients (MFCC) and spectrograms. These representations capture frequency and time information important for speech recognition. Trial set. This allows the model to learn from one layer, generalize to another set, and evaluate its performance in a separate layer. CNN model architecture for audio processing. This model should have layers, layers, and layers all together. Try different configurations to optimize performance. Monitor your model's performance during the validation process and make adjustments as needed. Analyze metrics such as accuracy, precision, and recall to measure the model's ability to recognize different sounds.

### III. Music recommendation algorithm

Create music library Use Python to access music website to store music files and music files. The music file is stored in Excel format. Based on seven different behaviors that can be defined by the cognitive model (anger, conflict, patience, happiness, sadness, security, safe, neutral), the accessed songs are analyzed and divided into a song library

Song analysis . The method of this article will be divided into two steps: searching for lyric data and song ideation.

The lyrics data mining steps [13] are as follows (Using Chinese songs as an example, English songs are also done in a similar way):

#### a. Word segmentation

This paper uses the open Chinese song word segmentation system Jieba as a segmentation tool.

b. Comparison with descriptive language This world makes use of vocabularies created by others; dictionaries such as CNKI and Sogou, for example. In the dictionary, words are divided into two groups with different meanings, that is, a group containing words with positive emotions ("happy", "happy", "warm", etc.) and a process. written to distinguish positive words, such as "fall", "disappointment", etc. It contains words with negative emotions, such as, this sentence contains emotional words, advanced words (such as "good"). "happy") and negative words, compare songs and see the results of the emotions in the song.

#### Song Sentiment Analysis

This article uses the SVM model for sentiment classification. The principle of SVM is below (schematic diagram is shown in Figure 3): We need to divide the black points and white points in the image. SVM will detect the boundary between the plane where the black area is and the plane where the white area is, such as the black line in the picture. Two rays are used to represent the black line and the distance between points in two planes. The goal of SVM is to find the black line with the largest distance  $\text{Max}(w)$ . This paper uses the LIBSVM (Chang et al., 2011) SVM toolkit developed by Professor Lin Zhiren of National Taiwan University to classify R expression.

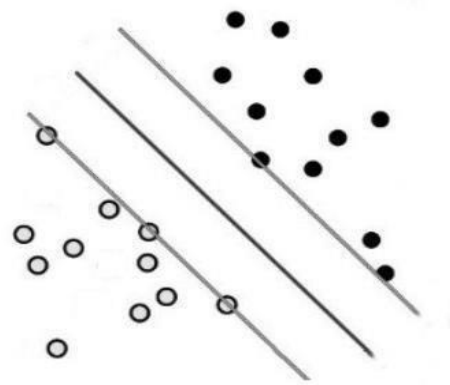


Figure 2: Schematic Diagram of SVM Classification

Since this article has seven different features, it needs a lot of classification. Each category is first divided into two categories (e.g., positive thoughts and negative thoughts), then the subcategories are divided into two subcategories, and so on until the items can be separated.

**Content-based recommendation algorithm (CB)** Content-based recommendation is to discover the relevance of items based on the metadata of the recommended items and then recommend similar products to the user based on the customer's previous preferences. According to the availability of the song library and the determination of facial expression, due to the different emotions expressed by different songs in the same mood, the life of the agreed proposal cannot play the role of reducing negative emotions. or encourage positive thinking. Thus, the initial recommendation algorithm is transformed into a content-based recommendation algorithm.

#### General steps for content-based recommendations General steps for recommendations are as follows:

Object representation: Extract some features of each object (this entire sentence is songs) to represent the object; = Education: Use specific information about various products that the user likes according to the user's preferences; Recommendation generation: By comparing the user's features with the candidate product obtained in the previous step, the product that is suitable for the user is recommended. The algorithm flowchart used in this article is shown in Figure 4:

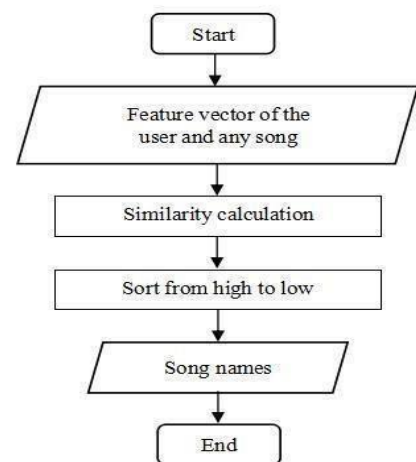


Figure 3: FlowChart of Content-based Recommendation Algorithm



### Feature Vector Extraction of Songs

In order to take full advantage of the data analysis of the pooling process, the model in this section combines global maximum and global average pooling to form a global pooling feature aggregation layer. The statistics between pooling and maximum pooling statistics are obtained by averaging the world and world maximum pooling of the custom map obtained from the GLU block layer. The results of the collaborative work here are one-dimensional. In Figure 5, two different rectangles with different colors are used to represent two one-dimensional features, and both statistical data are added to the next layer of the entire connection.

**IV. Experiment and Result Analysis** This experiment is carried out on the platform of TensorFlow deep learning framework, and the experimental data set is FER2013<sup>[15]</sup> public data set.

Figure 5 shows the loss function curve during training. The horizontal axis Epoch represents the number of training iterations, and the vertical axis Loss represents the average loss value of each group. Figure 6 shows the training recognition accuracy curve; The horizontal axis Epoch represents the number of training iterations, and the vertical axis Accuracy represents the batch.

It can be seen from Figure 8 that the model has the highest accuracy rate for identifying Happy and Surprise tags, and the accuracy rate is 83%; the model will have a deviation in identifying Disgust tags, and the probability of identifying as Angry is 51%. The probability of finding Disgust is only 29%; this model has the lowest recognition accuracy on Fear tags, only 23%. The reasons may be the following: Traditional Model . Need more video tutorials to improve learning style. Therefore, the macro expression dataset is included in the CK+ training data. Due to the limited number of frames on some CK+ models, some frames are printed to meet the 18 frame requirement for 3D input data. The macroscopic sentence will be different from the microscopic sentence. Therefore, in order to train our MER model using macro expression data, it is necessary to increase the similarity between macro expressions and micro expressions. That's why we use macro expression reduction. This algorithm assumes that the vertices of the ME are the same as the start of the macro expression and the expression between the vertices. According to this idea, the middle frame of the macro grid is selected as the corner frame of the image sequence for the training model, as shown in Figure 8. **Experimental Setup**

Due to the equal number of files in the original audio, each audio was converted to mono and down sampled at a 16 kHz sampling rate. The Fourier transform window length used when transforming the Mel spectrum is 512, the skip window size is 256, and the number of frequency bins is 128. Slicing time is 5 seconds with 50% overlap. Mel spectrum specification (313, 128) of a form produced after processing according to the above settings,

each sound pattern producing 11 parts of the same size.




Table 1		

Introduction to the GTZAN data set.

### Evaluation Index

Statistical accuracy (Acc) was chosen as the reference point for the music classification method discussed in this section. The accuracy of the distribution is calculated as follows:

$$\text{Acc} = \frac{N_C}{N} \times 100\%.$$

### V. Results:

Different models produce different results from different deep learning methods. For a fair comparison, all experiments are performed in the same environment and all parameters are kept to compare the proposed model with SVM, CNN, GLU, RCNN and RGLU. The laboratory has tested these networks with different models and the results are shown in Table 2 and Figure 6. Those used in the model are more conducive to acoustic spectral features than traditional convolutional learning. The gate model pays more attention to the audio spectrum features that are more important for the music classification task, allowing the features to pass to the next layer of the network, while data not relevant to the audio music classification function is ignored by the gate mechanism. Comparative experiments have confirmed the effectiveness of audio spectrum based gating models in music classification. From a data filtering perspective, GLU can be used as another implementation of the monitoring process. Unlike the RGLU model, which determines the weight of each channel specification, the GLU can be updated over time in the learning network. Surface weighting in one-dimensional convolution extends this time transition model, which increases the color in the time dimension, and together with one-dimensional convolution in the time dimension, it can give better results. Compared with CNN and GLU without residual models, the accuracy of RCNN and RGLU with residual models is improved; This suggests that the use of residual connections may improve classification accuracy for some reason. It is worth noting that the accuracy of RGLU using residual models is better than GLU, and the accuracy of RCNN is greater than CNN. This shows that the combination of residual sampling and gated convolution is suitable for data transfer in the network. Thus, this experiment proves all the benefits and efficiency of our algorithm. This paper presents a facial micro interpretation recognition model based on neural network (CNN). After training the model on the FER2013 dataset, we achieved a recognition rate of 62.1%. A content-based recommendation algorithm is used to determine music for the user through facial expression and emotion recognition. Compared with existing algorithms that recommend music based on the user's previous listening preferences, the algorithm proposed in this article makes the user's mood more familiar, ensuring that the

recommended music can be according to the user's listening needs. Therefore, this algorithm has a very large market. Although we have made some progress, there are still some issues to be resolved. For example, the accuracy of micro expression needs to be improved. In future studies, the recognition of tags with low recognition rate will be improved and the music recommendation algorithm will be improved.

## VI. Conclusion:

In this project, we created playlists based on user sentiment, developed an application to predict user sentiment using convolutional neural networks and used stream limit to create playlists. Let's try making music on the internet by opening Youtube. The method is to use a deep neural network (DNN) to learn the optimal hypothesis. DNN, object recognition, human prediction, face recognition, etc.

It is a method that has been successful in recent years. Convolutional neural networks (CNN) have proven useful in areas such as image recognition and classification. The proposed method can detect the user's facial expressions using the CNN model. When the needs are classified, songs suitable for the user's needs are played. In this project, the main website is created from the user's photos or recorded videos. The image/video is then sent to the server to predict the user's mood. Once the emotion is identified, the next step is to make music.

## VII. References:

- [1] Liu Jianwei, Liu Yuan, Luo Xionglin. Progress in Deep Learning Research [J]. Application Research of Computers, 2014, 31 (7): 1921-1942.
- [2] Shen Huijun. Research and implementation of face recognition image preprocessing method [J]. Science and Technology and Innovation, 2014 (18): 119-120.
- [3] Zhang Chen. Research on some key technologies of facial micro-expression recognition [D]. 2019.
- [4] Liu Mingqi, Ni Guoqiang, Chen Xiaomei. Research on Pretreatment Algorithm of Dorsal Vein Image [J]. Optics Technology, 2007, 33: 255-256.
- [5] Li Siqun, Zhang Xuanxiong. Research on Facial Expression Recognition Based on Convolutional Neural Networks [J]. Journal of Software, 2018, v.17; No.183 (01): 32-35.
- [6] Hou Yuqingyang, Quan Jicheng, Wang Hongwei. Overview of the development of deep learning [J]. Ship Electronic Engineering, 2017, 4: 5-9.
- [7] Liu Sijia, Chen Zhikun, Wang Fubin, et al. Multiangle face recognition based on convolutional neural network [J]. Journal of North China University of Technology (Natural Science Edition), 2019, 41 (4): 103- 108.
- [8] Li Huihui. Research on facial expression recognition based on cognitive machine learning [D]. Guangzhou: South China University of Technology, 2019.
- [9] Li Yong, Lin Xiaozhu, Jiang Mengying. Facial expression recognition based on cross-connection LeNet5 network [J]. Journal of Automation, 2018,44 (1): 176- 182. [10] Yao L S, Xu G M, Zhap F. Facial Expression Recognition Based on CNN Local Feature Fusion[J]. Laser and Optoelectronics Progress, 2020, 57(03): 032501.
- [11] Xie S, Hu H. Facial expression recognition with FRRCNN [J]. Electronics Letters, 2017, 53 (4): 235- 237.
- [12] Zou Jiancheng, Deng Hao. An automatic facial expression recognition method based on convolutional neural network [J]. Journal of North China University of Technology, 2019,31 (5): 51-56.
- [13] Xue Liang, Huang Meichuan. Emotional analysis of lyrics in Chinese popular music—Big data analysis method based on new media music terminal [D]. Music Culture Industry, 2017. (4): 77-81. \
- [14] Guo Yanhong. Research on collaborative filtering algorithm and application of recommendation system [D]. Dalian: Dalian University of Technology, 2008: 1- 41.
- [15] Mao Xu, Wei Cheng, Qian Zhao et al. Facial expression recognition based on transfer learning from deep convolutional networks[C]. 2015 11th Int. Conf. Nat. Comput.,2015: pp.702 – 708